Knowing or Guessing? Robust Medical Visual Question Answering via Joint Consistency and Contrastive Learning

Songtao Jiang[1,5], Yuxi Chen[1], Sibo Song[2], Yan Zhang[1], Yeying Jin[3], Yang Feng[4], Jian Wu[1,6], and Zuozhu Liu[1,6(✉)]

[1] Zhejiang University, Zhejiang, China
[2] Alibaba Group, Zhejiang, China
[3] National University of Sinapore, Sinapore
[4] Angelalign Technology Inc., Shanghai, China
[5] ChohoTech Inc., Hangzhou, China
[6] Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence, Zhejiang, China
zuozhuliu@intl.zju.edu.cn

Abstract. In high-stakes medical applications, consistent answering across diverse question phrasings is essential for reliable diagnosis. However, we reveal that current Medical Vision-Language Models (Med-VLMs) exhibit concerning fragility in Medical Visual Question Answering, as their answers fluctuate significantly when faced with semantically equivalent rephrasings of medical questions. We attribute this to two limitations: (1) insufficient alignment of medical concepts, leading to divergent reasoning patterns, and (2) hidden biases in training data that prioritize syntactic shortcuts over semantic understanding. To address these challenges, we construct RoMed, a dataset built upon original VQA datasets containing 144k questions with variations spanning word-level, sentence-level, and semantic-level perturbations. When evaluating state-of-the-art (SOTA) models like LLaVA-Med on RoMed, we observe alarming performance drops (e.g., a 40% decline in Recall) compared to original VQA benchmarks, exposing critical robustness gaps. To bridge this gap, we propose Consistency and Contrastive Learning (CCL), which integrates two key components: (1) knowledge-anchored consistency learning, aligning Med-VLMs with medical knowledge rather than shallow feature patterns, and (2) bias-aware contrastive learning, mitigating data-specific priors through discriminative representation refinement. CCL achieves SOTA performance on three popular VQA benchmarks and notably improves answer consistency by 50% on the challenging RoMed test set, demonstrating significantly enhanced robustness. Code will be released.

Keywords: Medical visual question answering · Medical vision-language models · Robustness and trustworthiness.

1 Introduction

Recent advancements in Medical Vision-Language Models (Med-VLMs), such as Med-Flamingo [23], Med-PaLM M [26], and LLaVA-Med [16], have demonstrated

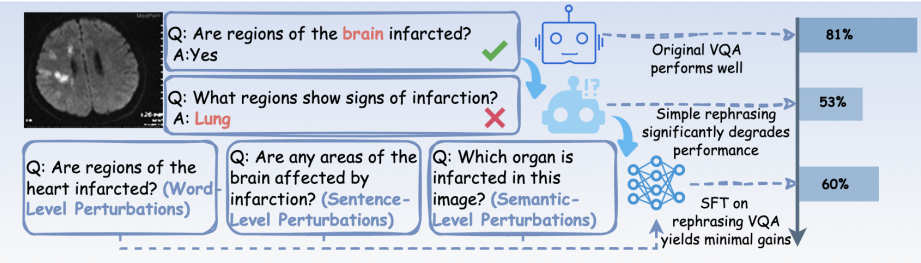

Fig. 1. A simple perturbation experiment demonstrates that current Med-VLMs exhibit inconsistencies in VQA tasks, raising concerns about the robustness of Med-VQA.

remarkable progress in Medical Visual Question Answering (Med-VQA) [13, 12, 11, 10, 21]. Through supervised fine-tuning (SFT) on Med-VQA training sets, these models achieve strong performance on downstream tasks. However, as illustrated in Fig. 1, our preliminary tests reveal a critical limitation: When questions are perturbed with varying levels of modifications while preserving semantic equivalence, models often produce inconsistent answers. This inconsistency severely restricts their applicability in real-world clinical settings, where diverse and interactive query formulations are common. Moreover, it raises fundamental concerns about current Med-VQA evaluation framework: *Is the model truly knowing the answers, or is it merely memorizing response patterns and guessing correctly by chance?*

To investigate further, we augmented the diversity of the original training data by introducing word-level perturbations. As shown in Fig. 1, vanilla SFT with more varied training data provides only marginal improvements in robustness against perturbations, with performance still significantly deviating from the original evaluation results. This yields two key insights: (1) the lack of diversity in training data contributes to the inconsistency issue [25], and increasing diversity can mildly mitigate it [24]; and (2) the current SFT paradigm, with its single autoregressive objective, has inherent limitations, as increasing data diversity alone provides minimal robustness gains. These findings highlight the need for a more robust Med-VQA evaluation framework and training methodology.

To address these challenges, we first construct the RoMed dataset as shown in Fig. 2, a new Med-VQA dataset encompassing training and testing sets across four major medical modalities: CT, MRI, X-Ray, and Pathology. For the training set, we enhance diversity by introducing multi-level perturbations at the word, sentence, and semantic levels, enriching the original Med-VQA training data. For the test set, we reconstruct a more comprehensive VQA benchmark based on mainstream Med-VQA datasets. Unlike traditional datasets [32, 30] that focus solely on accuracy, we incorporate evaluation metrics such as the Coefficient of Variation (CV) and Mean Absolute Deviation (MAD) to assess answer consistency, providing a more robust evaluation framework.

Furthermore, we propose Joint Consistency and Contrastive Learning (CCL) to address the limitations of the current SFT paradigm. Through consistency

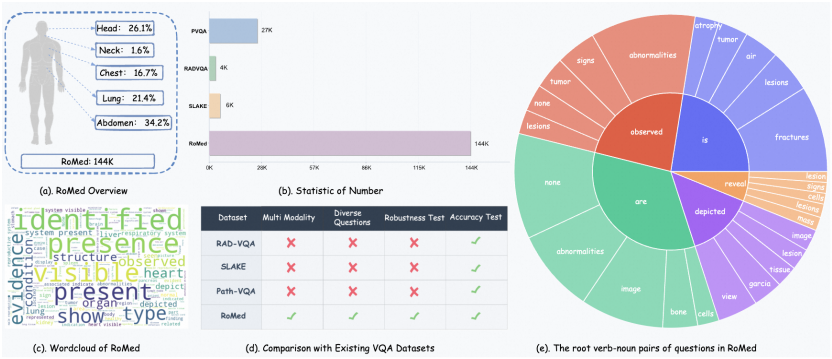

Fig. 2. Overview of RoMed. RoMed is a comprehensive VQA dataset spanning diverse organs and modalities (CT, MRI, X-Ray, Pathology), with dual evaluations for accuracy and robustness ensuring a holistic assessment.

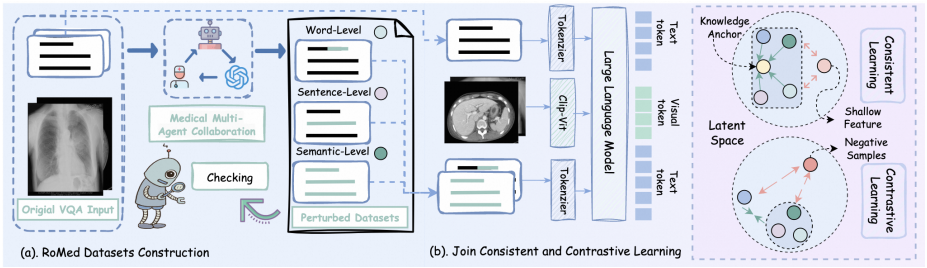

Fig. 3. Overview of our CCL pipeline. Our framework consists of two key components: (a) constructing the RoMed dataset through medical multi-agent collaboration; (b) joint knowledge-anchored consistency learning for medical expertise alignment and bias-aware contrastive learning to reduce inherent representation biases.

learning [6], CCL provides explicit supervision signals to ensure the model delivers correct answers across various perturbations, fostering better alignment with medical knowledge rather than shallow, overfitting features. Additionally, by treating perturbed questions as positive samples and using other questions in the same batch as negative samples, CCL guides the model to perform comparative understanding by leveraging contrastive learning objective [14]. This dual-objective approach mitigates potential overfitting in the model's representations and significantly enhances its generalization capabilities, making it more robust and reliable for real-world clinical applications. Extensive experiments and analyses demonstrate that CCL not only significantly enhances Med-VQA performance but also markedly reduces MAD and CV, thereby improving model robustness. CCL achieves state-of-the-art (SoTA) accuracy and robustness on widely-used benchmarks, including Rad-VQA [15], SLAKE [18], and PathVQA [8].

2 Methods

In this section, we first describe the construction of the RoMed dataset, addressing the lack of robustness evaluation in current Med-VQA systems. To tackle the limited generalization of vanilla SFT, we introduce our Joint Consistency and Contrastive Learning (CCL) framework, which optimizes Med-VLM representation learning by integrating consistency and contrastive objectives.

2.1 RoMed Datasets Construction

Our study reveals that current Med-VQA systems often fail to answer semantically equivalent perturbed questions correctly (see Fig. 1), despite accurately answering the original questions. This suggests that the reported accuracy on existing Med-VQA benchmarks may not reliably reflect the true knowledge level of Med-VLMs. To address this limitation, we construct a more diverse and robust evaluation dataset for Med-VQA (see Fig. 3 a). First, we integrate widely used Med-VQA datasets, including Rad-VQA [15], SLAKE [18], and PathVQA [8], which cover various organs and modalities. Building on these datasets, we introduce perturbations at three levels: word-level, sentence-level, and semantic-level, using a medical multi-agent collaboration system. Specifically, we leverage three models to generate high-quality perturbations, combining both general-purpose and domain-specific VLMs. For the medical multimodal agent, we employ HuatuoGPT-Vision-34B [5], a leading medical VLM, which provides domain-specific medical knowledge by generating captions for the given medical images. For the medical reasoning agent, we use HuatuoGPT-o1 [4], a single-modal LLM with advanced reasoning capabilities. This agent takes the captions and question-answer pairs as input to produce intermediate reasoning steps for sampling correct reasoning processes. Finally, we utilize GPT-4o as the general meta-agent, a state-of-the-art closed-source model, to integrate feedback from both the medical multimodal and reasoning agents, generating three levels of perturbed questions along with their corresponding answers. After this process, we validate the constructed questions by feeding them back to GPT-4o, ensuring they align with the same medical knowledge as the original questions and do not require additional knowledge beyond what is needed to answer the original questions correctly. Following this validation step, we construct the RoMed dataset, as shown in Fig. 2. Since perturbed questions are derived from the original ones, an ideal robust VLM should consistently answer all variants correctly within each question cluster. To quantify consistency, we introduce two metrics: Mean Absolute Deviation (MAD), defined as $\text{MAD} = \frac{1}{N}\sum_{i=1}^{N}|x_i - \mu|$, and Coefficient of Variation (CV), defined as $\text{CV} = \frac{\sigma}{\mu} \times 100\%$, where N is the number of questions in a cluster, x_i is the model's answer to the i-th question, μ is the mean of the answers, and σ is the standard deviation.

2.2 Joint Consistency and Contrastive Learning

Knowledge-Anchored Consistency Learning Let q denote the original question, tokenized into text tokens $\mathcal{T}_q$. The corresponding image I is encoded

into visual tokens $\mathcal{V}_I$ using a visual encoder (e.g., CLIP-ViT). The multimodal input $\mathcal{X}$ is formed by concatenating $\mathcal{T}_q$ and $\mathcal{V}_I$, i.e., $\mathcal{X} = [\mathcal{T}_q; \mathcal{V}_I]$. The input $\mathcal{X}$ is fed into the LLM backbone, generating outputs $\mathcal{Y}$. The autoregressive loss $\mathcal{L}_{\text{original}}$ is computed as: $\mathcal{L}_{\text{original}} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, \mathcal{X})$, where T is the output sequence length and y_t is the token at position t. To enhance the alignment with medical knowledge, we perform consistency learning by introducing perturbations at three levels: word-level, sentence-level, and semantic-level. These perturbations are constructed through multi-agent collaboration based on the original question. For each perturbed question q_i ($i \in \{w, s, \text{sem}\}$), the perturbed input $\mathcal{X}_i = [\mathcal{T}_{q_i}; \mathcal{V}_I]$ is used to compute the total consistency loss:

$$\mathcal{L}_{\text{consistency}} = \mathcal{L}_{\text{original}} + \sum_{i \in \{w,s,\text{sem}\}} \left(-\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, \mathcal{X}_i) \right).$$

Bias-Aware Contrastive Learning To eliminate bias in the training data and calibrate the model's representation, we employ contrastive learning as part of the CCL framework. Specifically, the original question q and its perturbed versions at three levels (q_w, q_s, and q_{sem}) are treated as positive samples, while other questions in the same batch serve as negative samples. The hidden state embedding $\mathcal{H}$ for the original input is obtained by feeding the multimodal input $\mathcal{X} = [\mathcal{T}_q; \mathcal{V}_I]$ into the LLM backbone and applying mean pooling: $\mathcal{H} = \mathcal{M}(\mathcal{LLM}(\mathcal{X}))$, where $\mathcal{M}(\cdot)$ denotes the mean pooling operation. Similarly, for each perturbed question q_p (with $p \in \{w, s, \text{sem}\}$), the corresponding hidden state embedding is $\mathcal{H}_p^+ = \mathcal{M}(\mathcal{LLM}([\mathcal{T}_{q_p}; \mathcal{V}_I]))$. The embeddings of unrelated questions in the batch are denoted as $\mathcal{H}_j^-$. The contrastive loss $\mathcal{L}_{\text{ctr}}$ is:

$$\mathcal{L}_{\text{ctr}} = -\sum_{p \in \{w,s,\text{sem}\}} \log \frac{\exp(\text{sim}(\mathcal{H}, \mathcal{H}_p^+)/\tau)}{\exp(\text{sim}(\mathcal{H}, \mathcal{H}_p^+)/\tau) + \sum_{j=1}^{N} \exp(\text{sim}(\mathcal{H}, \mathcal{H}_j^-)/\tau)},$$

where $\mathcal{H}_j^-$ represent the embedding of the j-th negative sample, $\text{sim}(\cdot, \cdot)$ denote the cosine similarity, $\tau > 0$ be a temperature hyperparameter, and N is the total number of negative samples. The overall loss is defined as $\mathcal{L} = \frac{\mathcal{L}_{\text{ctr}} + \mathcal{L}_{\text{consistency}}}{2}$.

3 Experiments

Evaluation Datasets and Metrics. To validate the effectiveness of our proposed CCL in enhancing traditional VQA performance, we conducted experiments on mainstream Med-VQA datasets, including Rad-VQA [15], SLAKE [18], and PathVQA [8]. These datasets cover CT, MRI, Chest-Xray, and Pathology modalities, encompassing both open-ended (free-form answers) and closed-ended (yes/no) question settings. For open-ended questions, we used Recall as the evaluation metric, while Accuracy was employed for closed-ended questions, consistent with prior research. Additionally, to assess the robustness of current Med-VLMs, we utilized our constructed RoMed dataset. Beyond Recall and Accuracy,

Method	RAD-VQA		SLAKE		PathVQA	
	Open	Closed	Open	Closed	Open	Closed
Representative & SoTA methods reported in the literature (Non-VLMs Based Methods)						
VL Encoder–Decoder [2]	-	82.5	-	-	-	85.6
Q2ATransformer [22]	-	81.2	-	-	54.9	88.9
Prefix T. Medical LM [27]	-	-	-	82.0	-	87.0
PubMedCLIP [7]	-	80.0	-	82.5	-	-
BiomedCLIP [31]	-	79.8	-	89.7	-	-
M2I2 [17]	-	83.5	-	91.10	-	88.0
BiomedGPT-S [30]	13.4	57.8	66.5	73.3	10.7	84.2
BiomedGPT-M [30]	53.6	65.0	78.3	86.8	12.5	85.7
CLIP-ViT w/ GPT2-XL	-	-	84.3	82.1	40.0	87.0
VLMs-based results						
GPT-4o [9]	51.6	64.0	59.1	71.6	24.1	76.0
LLaVA-v1.5 [19]	23.6	50.7	35.2	52.2	11.9	52.8
Med-Flamingo [23]	10.3	52.2	8.5	37.0	1.2	45.6
PMC-VQA [32]	6.3	41.5	7.3	33.9	1.0	40.1
SQ-LLaVA [29]	23.9	52.6	40.0	57.5	12.2	53.8
ST-LLaVA [28]	33.8	59.2	40.1	55.5	10.4	52.1
LLaVA-Med (StableLM)	51.6	75.4	82.2	82.7	33.2	89.5
LLaVA-Med (StableLM) + CCL	<u>62.7</u>	<u>84.9</u>	<u>83.6</u>	85.1	<u>36.3</u>	90.1
LLaVA-Med (Phi2)	54.5	79.8	82.1	<u>86.5</u>	34.0	<u>90.4</u>
LLaVA-Med (Phi2) + CCL	**65.0**	**88.2**	**83.8**	**88.5**	**37.5**	**90.7**

Table 1. Performance on traditional Med-VQA tasks. **Bold** denotes the best performance,<u>underlined</u> denotes the second-best.

Method	RoMed(RAD-VQA)				RoMed(SLAKE)				RoMed(PathVQA)			
	Recall	Acc	CV(↓)	MAD(↓)	Recall	Acc	CV(↓)	MAD(↓)	Recall	Acc	CV(↓)	MAD(↓)
LLaVA-Med (StableLM)	26.5	61.9	83.9	52.1	52.1	72.1	65.3	51.5	22.3	68.4	96.0	58.6
LLaVA-Med (StableLM) + CCL	<u>48.1</u>	<u>79.8</u>	<u>68.3</u>	<u>42.5</u>	<u>70.9</u>	<u>81.3</u>	<u>57.6</u>	<u>37.3</u>	<u>30.8</u>	<u>81.3</u>	<u>67.7</u>	<u>42.4</u>
LLaVA-Med (Phi2)	35.6	63.9	77.8	55.8	60.1	71.9	60.4	49.0	19.2	64.13	93.0	58.4
LLaVA-Med (Phi2) + CCL	**54.1**	**81.4**	**63.3**	**40.4**	**70.4**	**82.7**	**54.9**	**35.6**	**32.7**	**82.8**	**66.6**	**41.9**

Table 2. Performance on RoMed VQA. **Bold** denotes the best performance, <u>underlined</u> denotes the second-best. Note that lower values are better for CV and MAD.

we introduced MAD and CV coefficients to evaluate the consistency of reasoning, reflecting the robustness of Med-VLMs.

Implementation Details. For fair comparison, our hyperparameters align with LLaVA-Med [16]. We adopt pretrained CLIP-ViT-Large-Patch14 as the vision encoder and StableLM [3] and Phi2 [1] as LLM backbones. A 2-layer MLP is used as the projector, with training runs for 9 epochs with a learning rate of 2e-5 without weight decay and a batch size of 2, using 8 × RTX 3090 GPUs.

Baselines. We compare our method with several strong baselines: (1) *CLIP-based methods* (e.g., PubMedCLIP [7]), which achieve SOTA performance but are limited by their reliance on candidate words for open-ended questions; (2) *Medical foundation models* (e.g., BiomedGPT [30]), which leverage generative multimodal pretraining but lack multi-turn dialogue capabilities due to their non-LLM architecture; (3) *VLM-based models* (e.g., LLaVA-Med, LLaVA-v1.5 [16, 20]), which excel in VQA accuracy and interactive dialogue but prioritize precision over robustness. In contrast, our CCL method offers a plug-and-play enhancement for medical models, seamlessly integrating with VLM-based approaches to provide multi-turn dialogue support, improved accuracy, and enhanced robustness, making it ideal for real-world clinical applications.

A	B	RoMed-radvqa		RoMed-Slake		RoMed-Pvqa	
		Recall	Acc	Recall	Acc	Recall	Acc
x	x	35.6	63.9	60.1	71.9	19.2	64.1
x	✓	44.5	74.6	65.1	74.6	24.6	70.0
✓	x	40.7	65.0	62.3	73.5	20.3	64.9
✓	✓	**54.1**	**81.4**	**70.4**	**82.7**	**32.7**	**82.8**

Table 3. Ablation on joint learning. *A* denotes the consistency learning, and *B* denotes the contrastive learning.

Model	RoMed-radvqa		RoMed-Slake		RoMed-Pvqa	
	Recall	Acc	Recall	Acc	Recall	Acc
Baseline	35.6	63.9	60.1	71.9	19.2	64.1
CCL	54.1	81.4	70.4	82.7	**32.7**	82.8
CCL^{++}	**55.2**	**82.1**	**71.6**	**83.1**	32.4	**83.3**

Table 4. Model performance comparison under data scaling using LLaVA-Med (Phi2). The variant *CCL++* indicates training with doubled dataset size.

Traditional VQA Performance Comparison. As shown in Tab. 1, our CCL method, when integrated with the top-performing LLaVA-Med [16], achieves SOTA performance across three benchmarks. Notably, it excels in challenging open-ended questions, demonstrating its effectiveness as a plug-and-play module for robust VQA in clinical settings.

VQA Robustness Performance Comparison. We evaluated our approach on the RoMed VQA benchmark, which introduces variations to assess robustness under diverse clinical queries. As shown in Tab. 2, LLaVA-Med's accuracy drops significantly (e.g., RAD-VQA [15] recall decreases by nearly 50%). In contrast, with CCL, the model maintains high performance, reducing accuracy loss to within 20% (Fig. 4). This highlights the limitations of current VQA frameworks and underscores CCL's ability to enhance both performance and robustness for real-world applications.

3.1 Ablation and Analysis

Ablation of Joint Learning. We conducted experiments to validate the complementary roles of consistency learning and contrastive learning in our method. As shown in Tab. 3, the absence of either loss leads to a performance degradation. Contrastive learning plays a critical role in refining robust representations, while consistency learning ensures the model acquires knowledge across varied question formulations and establishes better alignment with medical knowledge. The combination of both components achieves the optimal performance.

Can SFT Improve VQA Robustness? To verify that our performance improvements are attributable to CCL rather than additional training data, we compared the performance of LLaVA-Med with CCL and vanilla SFT, both trained on the RoMed trainset. As shown in Fig. 5, vanilla SFT on a larger dataset fails to effectively enhance model robustness. This demonstrates the effectiveness of CCL, which leverages consistency learning to acquire new knowledge while utilizing contrastive learning to refine representations.

Representation Visualization Comparison. As shown in Fig. 4 (e) and (f), we observe that in vanilla LLaVA-Med, the embeddings of the three levels of variations and the original questions are widely separated, indicating that the representations fail to capture the shared features across different formulations. This sensitivity to perturbations could lead to misdiagnoses in real-world clinical applications with diverse query formulations. In contrast, with CCL, the model's

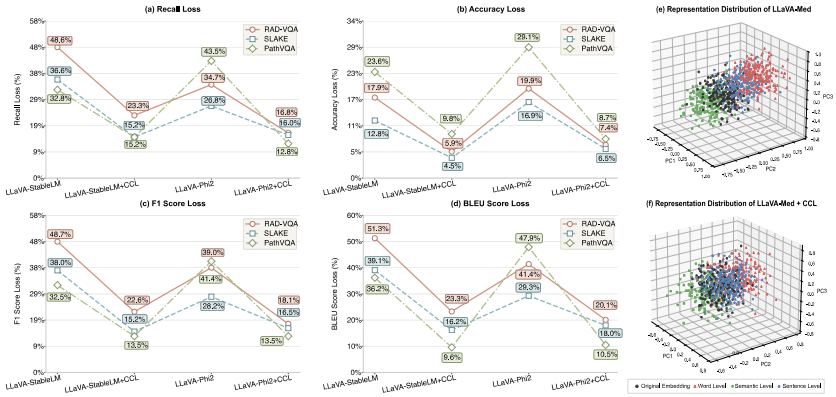

Fig. 4. (a)–(d) Performance degradation under varied VQA questions, significantly mitigated by CCL; (e)–(f) Representation embeddings of multi-level VQA variations.

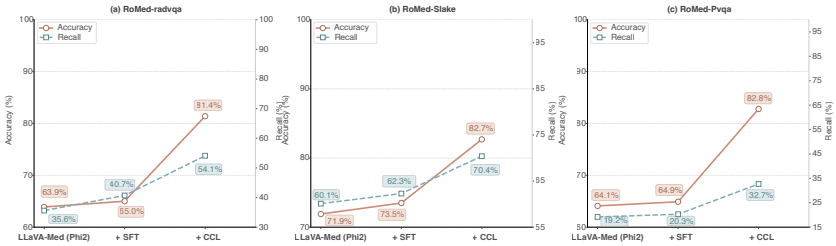

Fig. 5. Comparison between SFT and CCL. SFT yields minimal performance gains.

representations under varied perturbations become more robust, suggesting that the model learns more low-level, generalizable features across different levels of perturbations. This makes it better suited for high-stakes clinical scenarios.

Effect of Scaling Data. To evaluate the effectiveness of our method on larger-scale data, we expanded the original VQA questions by generating two additional variations per level (word-level, sentence-level, and semantic-level), resulting in a dataset twice the size of RoMed training data. This allowed us to explore the trade-off between performance and cost. As shown in Tab. 4, adding one variation per level significantly improves the model's VQA performance and robustness. However, doubling the dataset size yields only marginal gains. Considering the training time overhead, expanding by one variation per level enables the model to achieve strong generalization capabilities through CCL.

4 Conclusion

This work reveals the fragility of Med-VLMs in providing consistent answers to semantically equivalent medical questions, attributing it to insufficient concept alignment and training data biases. To address these challenges, we con-

struct RoMed, a dataset with diverse perturbations, and propose Consistency and Contrastive Learning (CCL), which enhances robustness by aligning models with medical knowledge and reducing biases, achieving state-of-the-art performance.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (Grant No. 12326612, 62476241), the Natural Science Foundation of Zhejiang Province, China (Grant No. LZ23F020008), and the Zhejiang University-Angelalign Inc. R&D Center for Intelligent Healthcare.

Disclosure of Interests. Yang Feng is employed by Angelalign Technology Inc.

References

1. Abdin, M., Jacobs, S.A., Awan, A.A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H., et al.: Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219 (2024)
2. Bazi, Y., Rahhal, M.M.A., Bashmal, L., Zuair, M.: Vision–language model for visual question answering in medical imagery. Bioengineering (2023)
3. Bellagente, M., Tow, J., Mahan, D., Phung, D., Zhuravinskyi, M., Adithyan, R., Baicoianu, J., Brooks, B., Cooper, N., Datta, A., et al.: Stable lm 2 1.6 b technical report. arXiv preprint arXiv:2402.17834 (2024)
4. Chen, J., Cai, Z., Ji, K., Wang, X., Liu, W., Wang, R., Hou, J., Wang, B.: Huatuogpt-o1, towards medical complex reasoning with llms (2024), https://arxiv.org/abs/2412.18925
5. Chen, J., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Zhang, R., Cai, Z., Ji, K., Yu, G., Wan, X., Wang, B.: Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale (2024), https://arxiv.org/abs/2406.19280
6. Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., Zisserman, A.: Temporal cycle-consistency learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
7. Eslami, S., Meinel, C., De Melo, G.: Pubmedclip: How much does clip benefit visual question answering in the medical domain? In: Findings of the Association for Computational Linguistics: EACL 2023. pp. 1151–1163 (2023)
8. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
9. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
10. Jiang, S., Wang, Y., Chen, R., Zhang, Y., Luo, R., Lei, B., Song, S., Feng, Y., Sun, J., Wu, J., et al.: Capo: Reinforcing consistent reasoning in medical decision-making. arXiv preprint arXiv:2506.12849 (2025)
11. Jiang, S., Wang, Y., Song, S., Zhang, Y., Meng, Z., Lei, B., Wu, J., Sun, J., Liu, Z.: Omniv-med: Scaling medical vision-language model for universal visual understanding. arXiv preprint arXiv:2504.14692 (2025)
12. Jiang, S., Zhang, Y., Jin, Y., Tang, Z., Wu, Y., Feng, Y., Wu, J., Liu, Z.: Hscr: Hierarchical self-contrastive rewarding for aligning medical vision language models. arXiv preprint arXiv:2506.00805 (2025)

13. Jiang, S., Zheng, T., Zhang, Y., Jin, Y., Yuan, L., Liu, Z.: Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 3843–3860 (2024)
14. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
15. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 1–10 (2018)
16. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36** (2024)
17. Li, P., Liu, G., Tan, L., Liao, J., Zhong, S.: Self-supervised vision-language pretraining for medical visual question answering. arXiv preprint arXiv:2211.13594 (2022)
18. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1650–1654. IEEE (2021)
19. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26296–26306 (June 2024)
20. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26296–26306 (2024)
21. Liu, J., Wang, Y., Du, J., Zhou, J.T., Liu, Z.: Medcot: Medical chain of thought via hierarchical expert. arXiv preprint arXiv:2412.13736 (2024)
22. Liu, Y., Wang, Z., Xu, D., Zhou, L.: Q2atransformer: Improving medical vqa via an answer querying decoder. arXiv preprint arXiv:2304.01611 (2023)
23. Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P., Rajpurkar, P.: Med-flamingo: a multimodal medical few-shot learner. In: Machine Learning for Health (ML4H). pp. 353–367. PMLR (2023)
24. Ray, A., Sikka, K., Divakaran, A., Lee, S., Burachas, G.: Sunny and dark outside?! improving answer consistency in vqa through entailed question generation. arXiv preprint arXiv:1909.04696 (2019)
25. Shah, M., Chen, X., Rohrbach, M., Parikh, D.: Cycle-consistency for robust visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6649–6658 (2019)
26. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al.: Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617 (2023)
27. van Sonsbeek, T., Derakhshani, M.M., Najdenkoska, I., Snoek, C.G., Worring, M.: Open-ended medical visual question answering through prefix tuning of language models. arXiv preprint arXiv:2303.05977 (2023)
28. Sun, G., Qin, C., Fu, H., Wang, L., Tao, Z.: Stllava-med: Self-training large language and vision assistant for medical question-answering. arXiv preprint arXiv:2406.19973 (2024)

29. Sun, G., Qin, C., Wang, J., Chen, Z., Xu, R., Tao, Z.: Sq-llava: Self-questioning for large vision-language assistant. In: European Conference on Computer Vision. pp. 156–172. Springer (2025)
30. Zhang, K., Yu, J., Yan, Z., Liu, Y., Adhikarla, E., Fu, S., Chen, X., Chen, C., Zhou, Y., Li, X., et al.: Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. arXiv preprint arXiv:2305.17100 (2023)
31. Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., et al.: Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv preprint arXiv:2303.00915 (2023)
32. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: Pmc-vqa: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415 (2023)