Med-CAP: Counterfactual Evidence and Adaptive Prior Suppression for Robust Medical Visual Question Answering

Zaiqiang Huang[1,2], Zhihong Zhu[2], Zixuan Huang[2], Yunyan Zhang[2], Hui Zhang[1], and Xian Wu[2⋆]

[1] Tsinghua University, Beijing, China
[2] Tencent Jarvis Lab, Beijing, China
zaiqhuang@gmail.com; kevinxwu@tencent.com

Abstract. Medical Visual Question Answering (Med-VQA) aims to provide clinical decision support by jointly reasoning over medical images and text. While existing research has made significant strides in addressing data scarcity through advanced representation learning and ensemble-based debiasing, current models still suffer from two critical issues: imbalance bias arising from skewed training distributions and modality shortcut bias where models rely on superficial linguistic patterns. In this paper, we propose Med-CAP, a counterfactual-and-calibration framework for robust Med-VQA. Specifically, to mitigate modality shortcuts, we introduce Counterfactual Evidence Modeling (CEM), which extracts prediction-relevant visual features via gradient attribution and enforces decision-making to be visually grounded through sufficiency-and-necessity constraints. To alleviate imbalance bias, we propose Adaptive Prior Suppression (APS), which calibrates fused logits by dynamically subtracting sample-adaptive language priors based on evidence strength. Experiments on SLAKE, VQA-RAD, SLAKE-CP and VQA-RAD-CP benchmarks show that Med-CAP achieves state-of-the-art robustness and accuracy across both standard and out-of-domain (OOD) settings. The code is available at https://github.com/ZaiqiangHuang/Med-CAP.

1 Introduction

Medical Visual Question Answering (Med-VQA) [13, 9] aims to answer clinical questions by jointly reasoning over medical images and text, which has gained increasing attention [2, 19, 8, 1]. Unlike general-domain models that benefit from massive data, Med-VQA models are prone to overfitting on limited samples. Therefore, models often generate predictions based on superficial associations rather than deep pathological reasoning, severely limiting their deployment in high-stakes clinical environments and motivating more robust approaches.

To address the data limitations in Med-VQA, existing research has achieved significant milestones in representation learning and bias mitigation [15]. For

⋆ Corresponding author

instance, methods like MEVF [20] have successfully introduced hybrid visual features to enhance image understanding when annotated data is scarce. Simultaneously, ensemble-based approaches, such as RUBi [5], have demonstrated the effectiveness of using a question-only branch to explicitly identify and subtract inherent linguistic biases from the joint predictions. Furthermore, the introduction of data re-weighting strategies has proven beneficial in balancing the model's attention across various answer categories [11, 7, 4, 22, 23].

Although these methods have achieved promising results, we still discover two critical issues that undermine the robustness of current Med-VQA systems. First, from the data perspective, the persistent imbalance bias acts as a root cause of language priors; models tend to over-rely on frequent answer distributions in the training set, which leads to catastrophic performance drops on rare but clinically vital cases [13, 9, 22, 23]. Second, from the reasoning perspective, models often succumb to modality shortcut bias, an effect where they bypass deep pathological visual evidence and instead rely on superficial linguistic patterns from standardized question templates [5, 11, 7, 4]. As a result, even state-of-the-art models often fail to generalize to out-of-distribution (OOD) clinical scenarios, where the visual evidence contradicts the historical language patterns [22, 23].

To address these challenges, we propose Med-CAP, a counterfactual-and-calibration framework for robust Med-VQA. Targeting the modality shortcut bias, we introduce Counterfactual Evidence Modeling (CEM). CEM extracts prediction-relevant visual evidence via feature-gradient attribution and constructs evidence-keeping and evidence-dropping counterfactual inputs. By enforcing a margin-based sufficiency-and-necessity constraint, CEM encourages the model to make decisions grounded in causally relevant visual evidence rather than superficial linguistic patterns. To alleviate the data imbalance bias, we propose Adaptive Prior Suppression (APS). APS calibrates the fused logits by subtracting a sample-adaptive question-only prior, where the suppression strength is dynamically determined by language prior confidence and counterfactual evidence strength. This dual-mechanism approach ensures that the model not only looks at the right visual regions but also remains resilient to skewed training distributions.

Our main contributions are summarized as follows: (1) We propose Med-CAP, a new Med-VQA framework that addresses medical language biases from both imbalance and shortcut perspectives. (2) We introduce CEM to ensure decisions are visually grounded via counterfactual constraints, and APS to calibrate predictions to alleviate dataset imbalance. (3) Experiments on standard (SLAKE, VQA-RAD) and bias-sensitive (SLAKE-CP, VQA-RAD-CP) [22] benchmarks demonstrate that Med-CAP achieves state-of-the-art robustness in OOD scenarios while maintaining competitive accuracy.

2 Methods

2.1 Preliminaries

We consider a medical VQA dataset $\mathcal{D} = \{(v^{(n)}, q^{(n)}, a^{(n)})\}_{n=1}^{N}$, where each sample consists of object-level visual features $v \in \mathbb{R}^{K \times d}$, a question represented

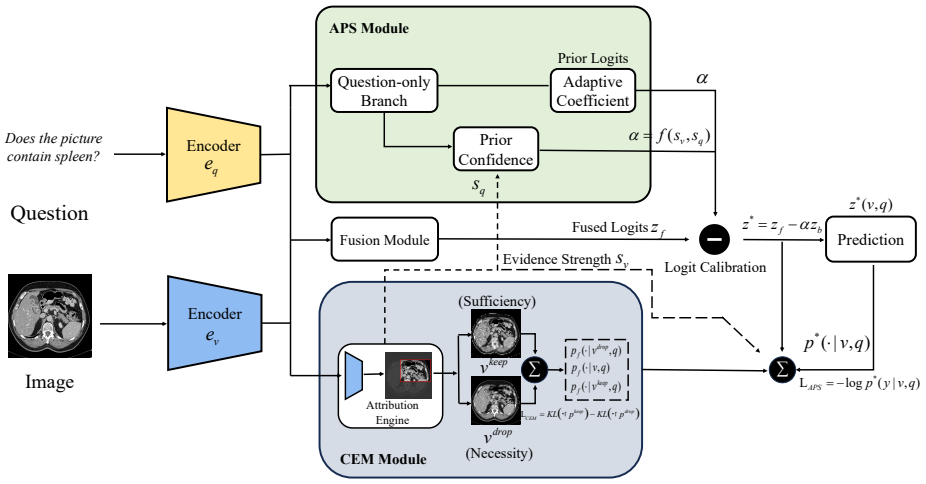

Fig. 1. Overview of the proposed Med-CAP framework for mitigating medical language biases. The framework consists of two complementary modules: Counterfactual Evidence Modeling (CEM) and Adaptive Prior Suppression (APS). The CEM module generates counterfactual feature representations to quantify the sufficiency and necessity of visual evidence, guiding the model to focus on clinically relevant cues rather than spurious correlations. The APS module dynamically adjusts prediction logits to suppress over-confident question-only priors, effectively mitigating dataset-specific biases.

as a token sequence $q \in \mathbb{N}^T$, and a one-hot encoded answer $a \in \{0,1\}^{|\mathcal{A}|}$. The goal is to learn a predictive model $f : \mathcal{V} \times \mathcal{Q} \to \mathbb{R}^{|\mathcal{A}|}$ that outputs logits corresponding to all candidate answers. However, existing Med-VQA models often suffer from medical language bias, where predictions can be dominated by question-only shortcuts rather than grounded visual evidence. To mitigate this issue, we introduce two plug-in modules, **Counterfactual Evidence Modeling (CEM)** and **Adaptive Prior Suppression (APS)**, which jointly promote evidence-grounded reasoning by enforcing counterfactual evidence consistency and suppressing over-confident question priors.

2.2 Counterfactual Evidence Modeling (CEM)

Motivated by counterfactual reasoning [16], we propose CEM to encourage causally grounded visual decision-making. The key idea is that if a prediction truly depends on visual evidence, then retaining such evidence should preserve the confidence of the correct answer, while suppressing it should reduce confidence.

Given a training sample with question q and ground-truth answer class $y \in \mathcal{A}$, the image is represented by K object-level features $v = \{v_k\}_{k=1}^{K}$, where $v_k \in \mathbb{R}^d$. Let z_y denote the fused logit corresponding to class y produced by the backbone. We estimate object-level importance via gradient-based attribution [18], for each object k, the attribution score is computed as the gradient magnitude of the

target logit w.r.t. the object feature, i.e., $s_k = \left\| \frac{\partial z_y}{\partial v_k} \right\|_2$, where $\|\cdot\|_2$ is the ℓ_2 norm. We normalize $\{s_k\}_{k=1}^K$ within each sample and select the top-ρ objects to obtain a binary evidence mask $M \in \{0,1\}^K$, where $M_k = 1$ indicates that object k is treated as evidence and $\rho \in (0,1)$ controls sparsity. Using M, we construct counterfactual features by suppressing visual signals with a scaling factor $\beta \in (0,1)$, i.e., $\tilde{v} = \beta v$. We then form evidence-keeping and evidence-dropping inputs:

$$v^{keep} = M \odot v + (1 - M) \odot \tilde{v}, \quad v^{drop} = (1 - M) \odot v + M \odot \tilde{v}, \tag{1}$$

where $\odot$ denotes element-wise multiplication (broadcast along the feature dimension). Intuitively, v^{keep} preserves evidence objects while attenuating non-evidence objects, whereas v^{drop} attenuates evidence objects and preserves the remaining context. Let $p^{keep}(\cdot) = p(\cdot|v^{keep}, q)$ and $p^{drop}(\cdot) = p(\cdot|v^{drop}, q)$ denote the predictive distributions after softmax normalization of the corresponding logits. We enforce a margin between their confidence on the correct class y:

$$\mathcal{L}_{CEM} = \max\Big(0,\ m - \big[\log p^{keep}(y) - \log p^{drop}(y)\big]\Big), \tag{2}$$

where $m > 0$ is a margin hyperparameter. Minimizing $\mathcal{L}_{CEM}$ promotes sufficiency via increasing $p^{keep}(y)$ and necessity via decreasing $p^{drop}(y)$.

2.3 Adaptive Prior Suppression (APS)

Although CEM encourages evidence-based reasoning, Med-VQA models may still exploit question-conditioned priors. Following the question-only prior paradigm in debiased VQA [5, 6], we introduce APS to suppress such priors in an instance-adaptive manner. Specifically, we add an auxiliary question-only branch that outputs logits $z_b(q) \in \mathbb{R}^{|\mathcal{A}|}$, from which we obtain the prior distribution $p_b(\cdot|q) = \mathrm{softmax}(z_b(q))$. We compute a prior-strength score s_q from $p_b(\cdot|q)$, where larger s_q indicates a more confident (lower-entropy and peakier) question-only prediction. To reflect the reliability of visual evidence, we compute an evidence-strength score using CEM counterfactual predictions as $s_v = p^{keep}(y) - p^{drop}(y)$, where larger s_v indicates that keeping evidence increases confidence for the correct answer. The sample-adaptive suppression coefficient is defined as:

$$\alpha(v, q) = \mathrm{clip}\big(\alpha_0 + \alpha_1 s_q - \alpha_2 s_v,\ \alpha_{\min}, \alpha_{\max}\big), \tag{3}$$

where α_0 is a base level, α_1 and α_2 control the sensitivity to s_q and s_v, and $\mathrm{clip}(\cdot, \alpha_{\min}, \alpha_{\max})$ constrains α to a fixed interval. Let $z_f(v, q) \in \mathbb{R}^{|\mathcal{A}|}$ denote the fused logits from the backbone. We calibrate them by subtracting the scaled prior logits, $z^*(v, q) = z_f(v, q) - \alpha(v, q)\, z_b(q)$ and $p^*(\cdot|v, q) = \mathrm{softmax}(z^*(v, q))$, then APS is optimized with cross-entropy on the calibrated distribution:

$$\mathcal{L}_{APS} = -\log p^*(y|v, q), \tag{4}$$

where $p^*(y|v, q)$ denotes the calibrated probability assigned to the golden class.

Table 1. Performance comparison on SLAKE-CP and VQA-RAD-CP datasets.

Methods	SLAKE-CP			VQA-RAD-CP		
	All	**Open**	**Closed**	**All**	**Open**	**Closed**
SAN [19]	26.02	48.30	6.42	16.29	59.73	6.05
MFB [20]	30.56	55.70	8.44	22.53	72.12	10.84
BAN [8]	17.50	30.90	5.72	17.30	67.70	5.42
UpDn [1]	31.45	59.90	6.42	26.67	74.78	15.33
MEVF+SAN [15]	18.62	32.60	6.33	22.11	68.14	11.26
MEVF+BAN [15]	19.33	35.00	5.54	19.07	62.39	8.86
RUBi [5]	33.88	60.30	10.64	81.27	60.62	86.13
LPF [11]	40.34	43.70	37.38	41.52	65.04	35.97
GGE-iter [7]	35.05	61.30	11.96	21.60	51.33	14.60
RMLVQA [4]	76.42	60.50	**90.41**	89.45	69.03	94.26
CEDO [22]	79.27	66.70	90.33	92.07	73.45	**96.45**
Med-BiasX [23]	78.33	64.80	90.24	91.48	76.11	95.10
Qwen2.5-VL-7B-Instruct [3]	70.85	61.60	78.98	65.63	53.50	75.30
InternVL3-8B [24]	70.43	62.50	77.40	65.19	50.50	76.89
Med-Flamingo [14]	30.65	43.15	16.42	45.68	36.00	53.39
HuatuoGPT-Vision-7B [21]	68.65	62.30	74.23	67.18	58.00	74.50
MedGemma-1.5-4B-IT [17]	65.56	65.40	65.70	62.08	45.50	75.30
LLaVA-Med-v1.5-Mistral-7B [10]	50.02	57.12	41.94	45.68	35.00	54.18
HealthGPT [12]	25.18	23.30	26.82	58.54	47.00	67.73
Med-CAP	**79.97**	**68.30**	90.24	**92.24**	**76.99**	95.83
Improvement ↑	+0.70%	+1.60%	-0.17%	+0.17%	+0.88%	-0.62%

2.4 Overall objective

For each training sample, the final loss is defined as

$$\mathcal{L} = \mathcal{L}_{base} + \mathcal{L}_{CEM} + \mathcal{L}_{APS}, \tag{5}$$

where $\mathcal{L}_{base}$ follows the training objective of RMLVQA [4], and the underlying Med-VQA backbone is instantiated with the UpDn architecture [1]. The overall framework is illustrated in Fig. 1, and we name the proposed method **Med-CAP**.

3 Experiments

3.1 Datasets and Implementation Details

We evaluate Med-CAP on two widely-used Med-VQA benchmarks, SLAKE [13] and VQA-RAD [9]. To assess robustness against language priors, we additionally adopt two bias-sensitive benchmarks, SLAKE-CP and VQA-RAD-CP [22]. Following previous works, we report accuracy (%) on *Open*, *Closed*, and *All* question types. We implement all models in PyTorch and train on a single NVIDIA A100 GPU. We adopt UpDn [1] as the backbone architecture and use AdamW with weight decay 1×10^{-3}. The batch size is set to $B = 60$ and the learning rate is 2×10^{-3}. All results are obtained by averaging the scores over three runs with different random seeds. Following the standard VQA evaluation protocol adopted by prior Med-VQA debiasing works [4, 22, 23], accuracy for each question type is computed over the entire test set; samples whose ground-truth answers are absent from the training answer vocabulary contribute zero to their type's accuracy.

Table 2. Performance comparison on SLAKE and VQA-RAD datasets.

Methods	SLAKE			VQA-RAD		
	All	**Open**	**Closed**	**All**	**Open**	**Closed**
SAN [19]	76.00	74.00	79.10	52.89	31.64	65.50
MFB [20]	73.89	71.63	77.40	54.10	41.90	62.13
BAN [8]	76.25	75.97	76.68	55.43	48.60	59.93
UpDn [1]	81.34	79.84	83.65	66.74	51.40	76.47
MEVF+SAN [15]	75.97	74.72	77.88	60.71	40.65	74.05
MEVF+BAN [15]	77.76	75.97	80.53	62.34	43.09	75.14
RUBi [5]	78.42	76.43	81.49	51.22	36.87	60.66
LPF [11]	75.59	73.33	79.09	56.32	49.72	60.66
GGE-iter [7]	79.83	79.22	80.77	65.19	49.16	75.74
RMLVQA [4]	81.43	80.47	82.93	65.41	49.16	76.10
CEDO [22]	83.41	81.09	87.02	67.41	**58.66**	73.16
Med-BiasX [23]	82.47	80.62	86.34	67.42	50.64	**78.46**
Qwen2.5-VL-7B-Instruct [3]	68.43	66.76	76.62	65.85	52.50	76.49
InternVL3-8B [24]	73.02	71.02	83.10	65.41	52.00	76.10
Med-Flamingo [14]	23.78	22.31	30.99	37.47	32.00	41.83
HuatuoGPT-Vision-7B [21]	66.92	63.88	71.63	67.63	58.00	75.30
MedGemma-1.5-4B-IT [17]	69.87	68.83	75.21	60.31	47.00	70.92
LLaVA-Med-v1.5-Mistral-7B [10]	51.43	50.72	54.93	44.12	32.50	53.39
HealthGPT [12]	26.12	24.72	32.96	32.37	17.50	44.22
Med-CAP	**84.26**	**81.65**	**88.43**	**69.18**	55.31	78.31
Improvement ↑	+0.85%	+0.56%	+1.41%	+1.76%	-3.35%	-0.15%

3.2 Comparison with State-of-the-Art Methods

We compare with: (i) classical VQA models (SAN, MFB, BAN, UpDn), (ii) medical debiasing baselines (MEVF-based), (iii) natural-image debiasing methods adapted to Med-VQA (RUBi, LPF, GGE-iter, RMLVQA, CEDO, Med-BiasX), and (iv) recent vision-language models (VLM/medical VLM) including Qwen2.5-VL, InternVL3, Med-Flamingo, HuatuoGPT-Vision, MedGemma, LLaVA-Med, and HealthGPT. All reported numbers are from the original papers or reproduced under the same protocol for fair and consistent comparison.

Table 1 summarizes results on SLAKE-CP and VQA-RAD-CP. Our method achieves the best overall performance on both datasets. Notably, on SLAKE-CP, our model improves the overall accuracy from 79.27 to 79.97 compared with the strongest competitor, and also yields consistent gains on Open questions. On VQA-RAD-CP, our method achieves 92.24 overall accuracy, demonstrating improved robustness to language priors. Table 2 reports results on the original SLAKE and VQA-RAD benchmarks. Our approach achieves the best overall performance, indicating that the proposed debiasing strategy improves robustness to bias without compromising standard predictive performance, thereby striking a favorable balance between bias mitigation and in-distribution accuracy.

3.3 Ablation Study

We conduct ablations to isolate the contribution of each component. Table 3 reports results on the bias-sensitive benchmarks (SLAKE-CP and VQA-RAD-

Table 3. Ablation study of CEM and APS on SLAKE-CP and VQA-RAD-CP.

Methods	CEM	APS	SLAKE-CP			VQA-RAD-CP		
			All	Open	Closed	All	Open	Closed
Baseline	–	–	76.42	64.80	**90.41**	89.54	73.89	93.22
w/ CEM	✓	–	78.15	65.00	89.71	91.39	72.12	**95.93**
w/ APS	–	✓	78.10	64.20	90.32	91.05	73.45	95.20
Med-CAP	✓	✓	**79.97**	**68.30**	90.24	**92.24**	**76.99**	95.83

Table 4. Ablation study of CEM and APS on SLAKE and VQA-RAD.

Methods	CEM	APS	SLAKE			VQA-RAD		
			All	Open	Closed	All	Open	Closed
Baseline	–	–	81.43	80.47	82.93	65.41	49.16	76.10
w/ CEM	✓	–	83.13	80.87	86.74	68.29	54.75	77.21
w/ APS	–	✓	82.28	79.15	87.10	67.85	54.19	76.84
Med-CAP	✓	✓	**84.26**	**81.65**	**88.43**	**69.18**	**55.31**	**78.31**

CP), while Table 4 reports results on the standard benchmarks (SLAKE and VQA-RAD). Adding either CEM or APS consistently improves the baseline, and combining both yields the best performance across datasets. These results validate that CEM provides effective counterfactual supervision, while APS further stabilizes predictions by suppressing question-only shortcuts when visual evidence is weak. These ablation results confirm that the proposed components highlight the effectiveness and generalizability of the full model design.

3.4 Parameter Analysis

We conduct a parameter analysis of Med-CAP on SLAKE-CP to study the influence of the key hyperparameters in CEM and APS. As shown in Fig. 2, Med-CAP achieves the best performance with $\beta = 0.2$ and $(\alpha_0, \alpha_1, \alpha_2) = (0.5, 0.3, 0.2)$. This suggests that a moderate counterfactual feature suppression in CEM, together with an appropriate adaptive prior suppression schedule in APS, is crucial for improving robustness and overall accuracy.

3.5 Case Analysis

To qualitatively assess Med-CAP, we compare it with the baseline on SLAKE-CP (Fig. 3). The baseline tends to predict the majority answers from the training distribution, indicating a reliance on language priors rather than visual evidence. For instance, it incorrectly follows dominant training biases (e.g., predicting frequent answers such as "2" or "Yes") despite opposite ground truths and shifted test distributions, and its attention regions are less aligned with clinically relevant cues. Med-CAP suppresses question-only priors and produces predictions grounded in relevant visual evidence. Consequently, it focuses on more appropriate regions and correctly answers both prior-opposed cases, demonstrating improved robustness to language bias under distribution shift. These qualitative

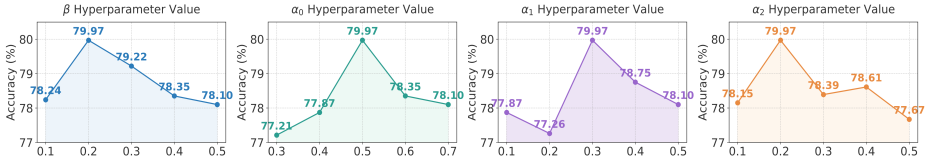

Fig. 2. Accuracy comparison on SLAKE-CP under different hyperparameter settings.

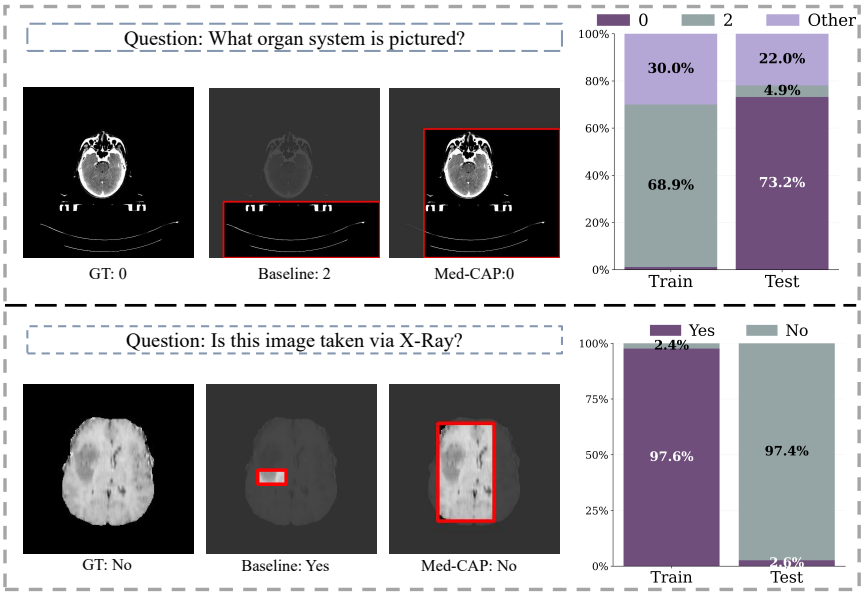

Fig. 3. Case analysis of the proposed Med-CAP. Two representative examples are shown for the "Yes/No" (up) and "How many" (down) question types.

results further corroborate the quantitative findings, highlighting that Med-CAP improves interpretability and robustness to language bias, particularly under distribution shift scenarios commonly encountered in medical VQA tasks.

4 Conclusion

In this paper, we propose Med-CAP, a counterfactual-and-calibration framework for robust medical visual question answering. By integrating Counterfactual Evidence Modeling (CEM) and Adaptive Prior Suppression (APS), our method promotes evidence-grounded reasoning while mitigating medical language bias. CEM enforces the sufficiency and necessity of visual evidence through counterfactual supervision, whereas APS adaptively suppresses over-confident question-only priors based on prior strength and evidence strength. Extensive experiments on both standard benchmarks (SLAKE/VQA-RAD) and bias-sensitive protocols

(SLAKE-CP/VQA-RAD-CP) demonstrate that Med-CAP improves robustness to language bias while maintaining competitive in-distribution accuracy. Future work will explore extending the framework to larger-scale medical vision-language models and more complex clinical scenarios.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6077–6086 (2018)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
4. Basu, A., Addepalli, S., Babu, R.V.: Rmlvqa: A margin loss approach for visual question answering with language biases. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11671–11680 (2023)
5. Cadene, R., Dancette, C., Cord, M., Parikh, D., et al.: Rubi: Reducing unimodal biases for visual question answering. Advances in neural information processing systems **32** (2019)
6. Clark, C., Yatskar, M., Zettlemoyer, L.: Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language Processing (EMNLP-IJCNLP). pp. 4069–4082 (2019)
7. Han, X., Wang, S., Su, C., Huang, Q., Tian, Q.: Greedy gradient ensemble for robust visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1584–1593 (2021)
8. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks. Advances in neural information processing systems **31** (2018)
9. 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), 180251 (2018)
10. 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**, 28541–28564 (2023)
11. Liang, Z., Hu, H., Zhu, J.: Lpf: A language-prior feedback objective function for de-biased visual question answering. In: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. pp. 1955–1959 (2021)

12. Lin, T., Zhang, W., Li, S., Yuan, Y., Yu, B., Li, H., He, W., Jiang, H., Li, M., xiaohui, S., Tang, S., Xiao, J., Lin, H., Zhuang, Y., Ooi, B.C.: HealthGPT: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. In: Forty-second International Conference on Machine Learning (2025), openReview: WbP2OwMULq
13. 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)
14. 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)
15. Nguyen, B.D., Do, T.T., Nguyen, B.X., Do, T., Tjiputra, E., Tran, Q.D.: Overcoming data limitation in medical visual question answering. In: International conference on medical image computing and computer-assisted intervention. pp. 522–530. Springer (2019)
16. Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.S., Wen, J.R.: Counterfactual vqa: A cause-effect look at language bias. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12700–12710 (2021)
17. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
18. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
19. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 21–29 (2016)
20. Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 1821–1830 (2017)
21. Zhang, H., Chen, J., Jiang, F., Yu, F., Chen, Z., Chen, G., Li, J., Wu, X., Zhiyi, Z., Xiao, Q., et al.: Huatuogpt, towards taming language model to be a doctor. In: Findings of the association for computational linguistics: EMNLP 2023. pp. 10859–10885 (2023)
22. Zhu, H., Liu, Y., Fang, X., Lu, G., Chen, B.: Cause-effect driven optimization for robust medical visual question answering with language biases. arXiv preprint arXiv:2506.17903 (2025)
23. Zhu, H., Liu, Y., Zhou, C., Lu, G., Chen, B.: Med-biasx: Robust medical visual question answering with language biases. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 369–378. Springer (2025)
24. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)