This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

I^2-Med: Interpretable Medical Inference through Visual-Guided Dynamic Logits Calibration

Md Sajid[1], Shreeyut Maheshwari[1], Akshat Mishra[1], Ritik Mishra[1], and Mohammad Tanveer[1]

Indian Institute of Technology Indore, Indore 453552, Madhya Pradesh, India
{phd2101241003, me230003071}@iiti.ac.in, akshatm430@gmail.com, {phd2301241003, mtanveer}@iiti.ac.in

Abstract. Medical vision language models have demonstrated strong potential for medical image understanding and reasoning. However, their interpretability and reliability remain limited because medical vision language tasks require precise visual grounding and clinically coherent explanations. This is particularly challenging due to complex imaging patterns and scarce expert annotations. These challenges reduce the effectiveness of conventional reasoning methods and lead to visual hallucination, defined as the generation of text that is not grounded in the given visual input, which further limits their applicability in clinical practice. To mitigate these challenges, we propose an interpretable medical inference (I^2-Med) framework that enhances visual faithfulness during inference without additional training to improve reasoning and explanation faithfulness in medical vision language models. During text generation, I^2-Med uses a visual guidance module to check whether each candidate word is supported by the medical image and consistent with the current reasoning. Based on this guidance, the model's output scores are adjusted to favor visually grounded words, without changing the model parameters. Extensive experiments on the OmniMedVQA benchmark, covering eight medical imaging modalities and five clinical question types, demonstrate that our proposed I^2-Med framework generates more clinically meaningful and visually faithful reasoning chains and reduces visual hallucination under standard medical VQA evaluation metrics. The source code is publicly available at `https://github.com/mtanveer1/l2-med`.

Keywords: Medical Vision Language Models · Interpretablity · Inference Time Calibration · BiomedCLIP · Visual Hallucination.

1 Introduction

The recent advancement of Large Multimodal Models (LMMs) [1] has shown a paradigm shift in medical artificial intelligence, promising transformative applications in automated diagnosis, report generation, and clinical decision support [13,8]. In comparison to traditional deep learning models restricted to specific modalities or tasks, medical Vision Language Models (VLMs) [12] possess the capacity to interpret complex visual data and articulate findings through natural

language. However, their deployment in clinical practice is hindered by two fundamental limitations: (i) the lack of transparent reasoning processes [2] and (ii) visual hallucination, the generation of plausible but image-ungrounded medical statements [14]. In high-stakes medical settings, where interpretability and factual accuracy are paramount, such failures can lead to dangerous misdiagnoses and an erosion of clinical trust.

Current approaches to adapt VLMs to the medical domain primarily rely on Supervised Fine-tuning (SFT) on medical image-text pairs [11]. While SFT enables models to learn domain-specific terminology, it often fails to instill robust reasoning capabilities. Models trained via SFT tend to overfit to the statistical correlations of the training data rather than learning to verify visual evidence, leading to superficial fluency without genuine understanding [3]. To address this, recent works have explored Reinforcement Learning (RL) strategies, such as Proximal Policy Optimization (PPO) and Group Relative Policy Optimization (GRPO), to align models with human preference and encourage Chain-of-Thought (CoT) reasoning [14,7]. While RL effectively incentivizes structured reasoning (e.g., generating step-by-step explanations), it operates primarily on the textual reasoning level. Crucially, these methods optimize reasoning structure but do not explicitly enforce token-level alignment between generated explanations and the underlying medical image during inference. Consequently, even models capable of complex reasoning may still hallucinate details that do not exist in the medical data (e.g., MRI and PET scans) to maintain the logical flow of their self-generated text [2].

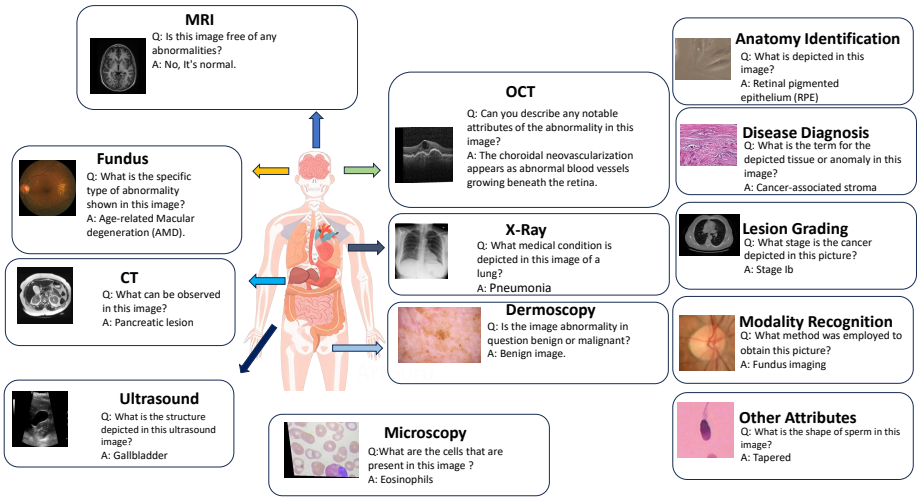

Fig. 1: Overview of the Evaluation Framework: Eight Medical Imaging Modalities and Five Medical Vision Question Answering Tasks.

To bridge this gap between logical reasoning and visual faithfulness, we propose I^2-Med, a novel unified framework for Interpretable Inference in medical VLMs. I^2-Med addresses the hallucination problem by two key components: (i) a lightweight Reinforcement Learning strategy to induce structured reasoning, and (ii) a training-free visual guidance mechanism based on Dynamic Logits Calibration (DLC) for faithfulness enforcement. Unlike standard decoding methods that rely solely on the language model's probability distribution, our approach introduces a visual sanity check at the decoding stage. By leveraging BiomedCLIP [15], a domain-specific contrastive vision language model, as an external visual alignment assessor, I^2-Med dynamically calibrates candidate token logits based on image-text consistency scores. This ensures that the generated text remains anchored to the visual evidence in real-time, penalizing tokens that drift from the image context without requiring computationally expensive retraining of the large backbone model. Our approach represents a significant step toward trustworthy medical AI by treating interpretability not as a post-hoc explanation, but as an intrinsic constraint of the generation process. The main contributions of this work are summarized as follows:

1. We introduce I^2-Med, a unified framework that integrates GRPO-based structured reasoning with inference-time visual-logit calibration to jointly enhance reasoning quality and visual faithfulness.
2. We propose MedCalib-IF, a training-free inference-time calibration mechanism that evaluates token-level visual grounding using BiomedCLIP and dynamically suppresses hallucinations through logits modulation.
3. We conduct extensive experiments on the OmniMedVQA benchmark, covering eight imaging modalities and various clinical tasks (as shown in Figure 1). Our results demonstrate that I^2-Med significantly reduces visual hallucination and improves the clinical validity of generated explanations compared to state-of-the-art baselines.

2 The proposed Methodology

We propose **I^2-Med**, a unified interpretable medical inference framework that enforces structured reasoning and visual faithfulness in medical VLMs. Unlike methods restricted to reinforcement learning or decoding-time correction, it employs a dual-stage design: (i) reward-driven structured reasoning during post-training and (ii) Visual-Guided Logits Regulation (VGLR) at inference, as demonstrated in Figure 2. To ensure clinical feasibility, we keep the pretrained medical VLM backbone fixed and avoid parameter updates at inference. Unlike existing approaches that require expensive retraining or fine-tuning, our method does not train the underlying VLM. Instead, interpretability is achieved exclusively during inference through policy optimization and decoding-time visual alignment. By leaving the pretrained model completely unchanged, our framework remains model-agnostic, computationally efficient, and readily deployable with any off-the-shelf VLM, enabling interpretable adaptation to medical tasks without additional training.

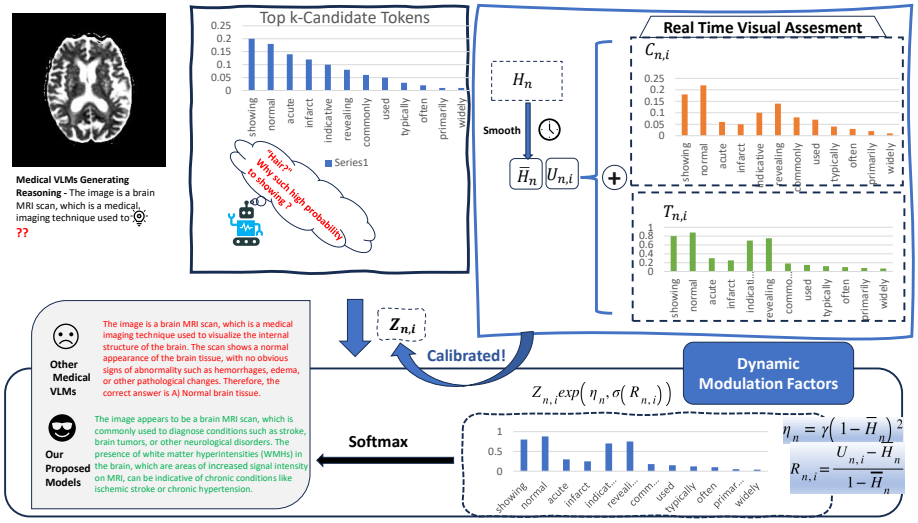

Fig. 2: Overview of Proposed I^2-Med Model.

2.1 Structured Policy Optimization for Medical Reasoning

To promote structured, clinically coherent reasoning without costly chain-of-thought annotations, we adopt a group-relative policy optimization strategy inspired by reinforcement learning–based alignment. Unlike supervised fine-tuning, which maximizes target likelihood, this method uses reward-driven updates to encourage diverse reasoning paths and structured intermediate explanations.

Optimization Objective of GRPO: Here, we discuss the optimization problem of GRPO, which we directly use in our proposed method. Let $\mathcal{D}$ denote the training dataset and $d \sim \mathcal{D}$ a sampled medical question. Let p_{old} and p_{new} represent the previous and updated policies, and p_{ref} a fixed reference policy from the pretrained backbone. For each question, we sample K candidate responses $r_i i = 1^K$ from $p_{\text{old}}(r \mid d)$.

The structured policy optimization objective is defined as:

$$\mathcal{L}_{\text{SPO}} = \mathbb{E}_{d\sim\mathcal{D},\, r_i\sim p_{\text{old}}} \left[\frac{1}{K} \sum_{i=1}^{K} \Big(\min\left(\rho_i A_i, \text{clip}(\rho_i, 1-\tau, 1+\tau) A_i\right) - \lambda\, D_{\text{KL}}(p_{\text{new}} \,\|\, p_{\text{ref}}) \Big) \right],$$

where $\rho_i = \frac{p_{\text{new}}(r_i|d)}{p_{\text{old}}(r_i|d)}$ denotes the policy ratio, τ is the clipping threshold, and A_i represents the group-normalized advantage.

Instead of using a learned critic, we compute A_i by normalizing rewards within each sampled response group. This group-relative approach improves training stability and reduces computational cost while enabling effective policy

updates. A KL divergence term regularizes learning, limiting deviation from the pretrained reference model and preserving multimodal knowledge.

Reward Formulation To ensure interpretable and diagnostically meaningful outputs, we employ a two-component hierarchical rule-based reward:

(1) Structural Consistency Reward. The model must place intermediate reasoning within `<think>` tags and the final decision within `<answer>` tags. A reward of 1 is assigned when this format is correctly followed, promoting organized and transparent reasoning.

(2) Diagnostic Accuracy Reward. A reward of 1 is given when the predicted answer matches the ground-truth multiple-choice label. A response is considered correct if the first token inside the `<answer>` tag corresponds to the correct option (e.g., "A", "B", "C", or "D").

The total reward is the sum of structural and accuracy components, ensuring both interpretability and diagnostic correctness.

2.2 Medical Visual-Guided Logits Regulation

To mitigate visual hallucination and ensure clinical faithfulness, we introduce *Medical Visual-Guided Logits Regulation (MVLR)*. MVLR employs BioMedCLIP as a domain-specific alignment model to enforce medically grounded consistency between radiological images and generated text. Operating entirely at inference, it requires no additional training or parameter updates and only one forward pass per decoding step, improving clinical consistency without added computational cost.

Contextual Medical Alignment Estimation: At decoding step n, prior to selecting token w_n, we evaluate the visual consistency of candidate tokens relative to the input medical image x. This process begins by estimating the visual grounding of the recent textual context.

Contextual Medical Alignment (CMA): We measure the alignment between the input image x and a fixed window of the most recent M generated tokens $w_{n-M:n-1}$ using BioMedCLIP: $S_n^{\text{CMA}} = \text{BioMedCLIP}(x, w_{n-M:n-1})$. This score quantifies how well the current reasoning context remains grounded in the visual evidence. To stabilize short-term fluctuations, we maintain a smoothed historical baseline $\bar{S}_n = \frac{1}{M}\sum_{j=n-M}^{n-1} S_j^{\text{CMA}}$, which represents the established level of visual grounding in recent generation steps.

Candidate-Level Visual Assessment: For the top-k candidate tokens $\{u_1, \ldots, u_k\}$ proposed by the medical VLM at step n, we compute two complementary alignment measures.

1. **Conditional Clinical Consistency (CCC):** This measures visual alignment when appending a candidate token to the current context: $S_{n,i}^{\text{CCC}} = \text{BioMedCLIP}(x, w_{n-M:n-1} \oplus u_i)$, where $\oplus$ denotes concatenation.
2. **Token Visual Affinity (TVA):** This measures the intrinsic alignment between the image and the candidate token: $S_{n,i}^{\text{TVA}} = \text{BioMedCLIP}(x, u_i)$.

Table 1: Comparison of performance across eight medical modalities against baseline models.

Methods \ Modality	CT	MRI	X-Ray	US	DER	FP	OCT	MICRO	Overall
Zero-shot VLMs									
BLIP-2† [9]	56.74	41.32	67.58	37.27	40.65	46.24	68.08	50.40	51.04
InstructBLIP† [4]	28.72	33.15	61.04	41.25	62.22	50.31	42.59	46.29	45.70
LLaVA† [10]	17.73	26.72	30.70	18.66	49.74	47.11	33.73	28.87	31.66
LLaMA Adapter v2† [5]	21.41	26.63	46.44	34.05	51.76	50.74	33.00	38.66	37.83
MiniGPT-4† [30]	22.81	27.48	38.30	25.50	40.25	38.33	31.40	36.23	32.54
InternVL2 [31]	40.20	58.10	57.90	49.10	51.90	53.20	59.10	64.00	54.19
Qwen2-VL-2B [3]	45.10	38.57	39.32	30.86	35.83	43.17	35.14	36.85	38.11
Qwen2-VL-7B [3]	61.46	45.77	64.27	36.01	49.08	59.84	59.32	61.08	54.60
Qwen2-VL-72B [32]	67.97	69.39	77.21	51.39	65.31	72.58	72.76	67.83	68.05
Qwen2.5-VL-3B [33]	53.87	54.23	61.84	32.69	52.94	62.47	56.23	59.64	54.24
Qwen2.5-VL-7B [33]	60.44	58.44	73.99	30.66	62.48	67.30	61.20	67.84	60.29
Qwen2.5-VL-72B [33]	66.18	68.74	77.59	49.81	69.75	71.04	69.22	69.37	67.71
Zero-shot Medical VLMs									
LLaVA-Med† [8]	18.69	27.47	30.68	29.88	44.95	39.03	34.61	33.29	32.33
RadFM† [34]	27.56	24.06	30.95	16.57	39.21	36.89	32.80	27.97	29.50
Med-Flamingo† [6]	38.47	40.56	30.34	24.64	32.43	30.12	26.51	19.93	30.38
MedVInT† [7]	40.74	43.10	55.10	41.26	29.11	31.84	23.26	32.00	37.05
HuatuoGPT-Vision [9]	35.30	40.40	41.50	60.10	53.10	51.40	59.30	62.30	50.43
HealthGPT [35]	35.50	78.50	81.90	51.40	64.90	54.60	89.30	88.20	68.04
Fine-tuned VLMs									
Qwen2-VL-2B (SFT)	51.74	52.83	65.57	47.65	51.91	52.26	53.99	56.58	54.07
Qwen2.5-VL-3B (SFT)	56.06	60.81	69.23	41.77	60.11	69.19	63.95	65.66	60.85
Qwen2-VL-2B (Think)	66.30	71.67	77.73	57.31	72.33	71.20	71.96	70.80	69.91
Qwen2-VL-2B (Think) (256 Token)	58.16	64.98	70.48	50.68	63.69	66.2	60.96	65.64	62.60
Proposed	60.21	65.29	71.10	51.75	66.42	69.32	63.79	66.41	65.29

Since contextual and intrinsic alignments capture complementary aspects of visual grounding, we compute a unified medical grounding score:

$$G_{n,i} = \frac{S_{n,i}^{\mathrm{CCC}} + S_{n,i}^{\mathrm{TVA}}}{2}. \tag{1}$$

This score provides a robust estimate of each candidate's clinical consistency.

Adaptive Clinical Logit Regulation Having computed the contextual baseline $\bar{S}_n$ and candidate grounding scores $G_{n,i}$, we adaptively regulate the original logits $Z_{n,i}$ produced by the VLM to favor clinically grounded tokens.

Relative Clinical Grounding (RCG) We quantify how much each candidate improves or degrades visual alignment relative to the recent baseline: $R_{n,i} = \frac{G_{n,i} - \bar{S}_n}{1 - \bar{S}_n}$. A positive value indicates stronger grounding than the recent contextual baseline.

Adaptive Clinical Constraint (ACC). The strength of intervention should depend on the current state of grounding. When the model is already well aligned with the image, aggressive regulation is unnecessary. Conversely, stronger correction is required when grounding weakens. We therefore define: $\alpha_n = \gamma(1 - \bar{S}_n)^2$, where γ controls the maximum modulation intensity.

Logit Update Rule. Finally, the calibrated logit is computed as: $Z'_{n,i} = Z_{n,i} \cdot \exp(\alpha_n \cdot \sigma(R_{n,i}))$, where $\sigma(\cdot)$ denotes the sigmoid function.

This multiplicative regulation increases the probability of clinically consistent tokens while suppressing visually unsupported ones. By incorporating domain-specific visual grounding into decoding, MVLR reduces semantic drift and hallucination without affecting fluency or inference efficiency.

Table 2: Comparison of our proposed models with medical VLMs baselines on five medical VQA tasks across five clinical reasoning types, evaluated under general-purpose zero-shot, medical zero-shot, and fine-tuned models.

Methods \ Types	Anatomy Identification	Disease Diagnosis	Lesion Grading	Modality Recognition	Other Attributes	Overall
Zero-shot VLMs						
BLIP-2	44.39	44.51	29.03	68.19	67.95	48.12
InstructBLIP	44.35	32.29	59.25	75.27	23.72	40.40
LLaVA	25.86	29.10	43.95	21.36	31.90	27.96
LLaMA Adapter v2	33.72	31.19	41.99	37.29	34.22	32.82
MiniGPT-4	28.88	30.47	34.56	26.43	30.36	29.74
Qwen2-VL-2B	30.70	36.53	43.58	59.90	42.19	42.58
Qwen2-VL-7B	42.57	48.83	52.06	84.74	59.66	57.57
Qwen2-VL-72B	56.41	65.71	62.15	98.11	80.53	72.58
Qwen2.5-VL-3B	35.23	50.45	52.79	85.23	54.77	55.69
Qwen2.5-VL-7B	41.00	61.32	54.13	97.78	70.45	64.94
Qwen2.5-VL-72B	57.22	62.55	60.32	98.20	77.41	71.14
Zero-shot Medical VLMs						
LLaVA-Med	29.53	29.22	34.18	26.93	33.08	29.25
RadFM	13.31	21.69	30.35	26.64	43.85	26.99
Med-Flamingo	24.93	38.90	30.74	30.19	14.18	34.03
MedVInT	40.26	35.78	12.77	68.10	30.30	40.04
Fine-tuned VLMs						
Qwen2-VL-2B (SFT)	53.97	51.62	60.71	86.77	63.91	63.39
Qwen2.5-VL-3B (SFT)	54.91	57.75	64.04	84.95	71.56	66.64
Qwen2-VL-2B (Think)	62.88	66.08	65.87	98.24	80.14	74.64
Qwen2-VL-2B (Think) (256 Token)	55.35	62.61	66.47	88.29	72.78	69.10
Proposed	56.00	64.03	66.43	88.96	73.58	69.80

3 Experiments

This section presents the datasets and results. We use the open-access dataset OmniMedVQA benchmark [6], a large-scale medical VQA dataset, for evaluation. Further implementation details are provided on GitHub.

3.1 Results Across Eight Medical Imaging Modalities

We evaluate I^2-Med across eight imaging modalities: CT, MRI, X-ray, Ultrasound, Dermoscopy, Fundus Photography, OCT, and Microscopy. Results on these eight modalities, with baseline comparisons are reported in Table 1. For controlled and efficient evaluation, reasoning length is limited to 256 tokens. We compare against three baseline groups: general-purpose zero-shot VLMs, zero-shot medical VLMs, and fine-tuned VLMs. I^2-Med consistently outperforms both zero-shot categories and achieves competitive, often superior, performance relative to fine-tuned models. Although Qwen2-VL-2B (Think) achieves slightly higher accuracy with unrestricted reasoning, it relies on substantially longer outputs and greater computational cost. Under the 256-token constraint, I^2-Med surpasses Qwen2-VL-2B (Think) (256 Token) in overall accuracy while maintaining structured, clinically grounded reasoning. Some cross-modality results have been discussed on our GitHub. These results demonstrate strong cross-modality generalization with improved efficiency, indicating that structured reasoning with medical visual grounding is more effective than merely increasing generation length.

3.2 Results Across Five Clinical Task Categories

We evaluate I^2-Med on five clinical tasks: Anatomy Identification, Disease Diagnosis, Lesion Grading, Modality Recognition, and Other Biological Attributes.

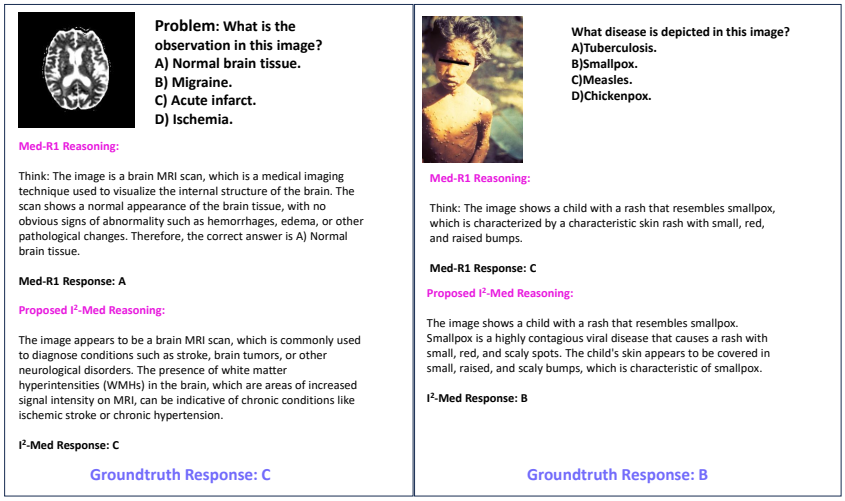

Fig. 3: Comparison of Med-R1 and I^2-Med on medical VQA examples shows that while both generate explicit reasoning, I^2-Med provides more clinically grounded explanations and achieves more accurate predictions.

The results are presented in Table 2. It achieves stable, competitive results, surpassing Qwen2-VL-2B (Think) (256 Token) in overall average under the 256-token limit. The slight 0.04 difference in Lesion Grading does not affect overall consistency. Detailed cross-modality results for these five clinical task categories are provided in our GitHub.

3.3 Interpretability: Qualitative Analysis of Generated Reasoning on Medical VQA

In Figure 3, we qualitatively compare I^2-Med with the Med-R1 baseline. In the first example (left), Med-R1 correctly identifies the modality but misses subtle pathological cues, leading to a false negative. In contrast, I^2-Med detects white matter hyperintensities and localizes the acute infarct, demonstrating more accurate and clinically grounded reasoning. In the second example (right), Med-R1 recognizes smallpox-related features but incorrectly predicts "Measles". I^2-Med maintains logical consistency, correctly associating umbilicated pustules with Smallpox. This shows that I^2Med effectively leverages its reasoning steps to reach the correct conclusion, whereas Med-R1 fails to translate its reasoning into an accurate prediction. Additional images related to qualitative analysis and post hoc interpretability are available on GitHub.

4 Conclusion

In this work, we proposed I^2-Med, an interpretable medical inference framework that enhances visual faithfulness in medical vision-language models through inference-time visual-guided logits calibration. By integrating structured policy optimization with dynamic token-level visual grounding, our method effectively reduces visual hallucination without requiring additional training or parameter updates. Extensive evaluations on the OmniMedVQA benchmark across eight imaging modalities and five clinical tasks demonstrate improved reasoning quality, clinical consistency, and competitive accuracy. Qualitative analyses further confirm that I^2-Med generates more diagnostically grounded and coherent explanations. Overall, our framework advances trustworthy and clinically deployable medical VLMs. Future work will investigate the effectiveness of our approach on a wider variety of medical VLM benchmarks and real-world clinical applications.

Acknowledgments. The work of Md Sajid is supported by the Council of Scientific and Industrial Research (CSIR), New Delhi, for providing fellowship under the Grant 09/1022(13847)/2022-EMR-I. This work is supported by PraxiaTech Private Limited. Further, this work is supported by IITI DRISHTI CPS Foundation under the National Mission on Interdisciplinary Cyber Physical System (NM-ICPS) of the Department of Science and Technology, Government of India.

Disclosure of Interests. The authors have no competing interests.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
2. Chen, J., He, J., Shao, Q., Chen, Q., Ying, J., Xu, H., Chen, J., Zheng, J., Wu, J.: Mitigating hallucination of large vision-language models via dynamic logits calibration. arXiv e-prints pp. arXiv–2506 (2025)
3. Chu, T., Zhai, Y., Yang, J., Tong, S., Xie, S., Schuurmans, D., Le, Q.V., Levine, S., Ma, Y.: Sft memorizes, rl generalizes: A comparative study of foundation model post-training. In: International Conference on Machine Learning. pp. 10818–10838. PMLR (2025)
4. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.H.: Instructblip: Towards general-purpose vision-language models with instruction tuning. In: Advances in Neural Information Processing Systems. vol. 36, pp. 49250–49267. Curran Associates, Inc. (2023)
5. Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., Li, H., Qiao, Y.: Llama-adapter v2: Parameter-efficient visual instruction model. arXiv preprint arXiv:2304.15010 (2023)
6. Hu, Y., Li, T., Lu, Q., Shao, W., He, J., Qiao, Y., Luo, P.: Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22170–22183 (2024)

7. Lai, Y., Zhong, J., Li, M., Zhao, S., Li, Y., Psounis, K., Yang, X.: Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. IEEE Transactions on Medical Imaging (2026)
8. 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)
9. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 19730–19742. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/li23q.html
10. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 34892–34916. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf
11. 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)
12. Pan, J., Liu, C., Wu, J., Liu, F., Zhu, J., Li, H.B., Chen, C., Ouyang, C., Rueckert, D.: 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. pp. 337–347. Springer (2025)
13. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
14. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y.K., Wu, Y., Guo, D.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models (2024), https://arxiv.org/abs/2402.03300
15. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)