EvoMed: Self-Evolving Medical VLM via Training-free Continued Learning

Jiaming Li[1†], Honglin Xiong[1†], and Qian Wang[1,2(✉)]

[1] School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai, China
qianwang@shanghaitech.edu.cn
[2] Shanghai Clinical Research and Trial Center, Shanghai, China

Abstract. Large vision-language models (VLMs) show strong general reasoning capabilities but remain unreliable for medical visual question answering. Existing medical VLMs often rely on supervised fine-tuning, which is limited by data scarcity, computational cost, and the risk of parameter-update-induced forgetting. We propose EvoMed, a training-free framework that enhances medical competency without updating model parameters. EvoMed employs a closed-loop self-evolving mechanism to iteratively distill clinical priors into a compact external knowledge base via reasoning, critique, and contrastive summarization. A coarse-to-fine retrieval mechanism selects highly relevant priors to guide inference toward clinically grounded predictions. Experiments on OmniMedVQA and SLAKE show that EvoMed improves both Qwen2.5-VL-7B and Qwen3-VL-235B-A22B by up to 7.5% using only 500 evolution samples. Code is available at https://github.com/Hunan-Tiger/EvoMed.

Keywords: Med-VQA · Training-free Adaptation · Retrieval-Augmented Generation

1 Introduction

Medical Visual Question Answering (Med-VQA) is an essential capability for clinical decision support, requiring AI systems to jointly interpret medical images and answer diagnostically meaningful questions [13]. Unlike general-domain VQA, Med-VQA demands the recognition of subtle, domain-specific visual cues and the application of specialized clinical knowledge, making it a substantially more challenging multimodal reasoning task. Despite remarkable progress in Vision-Language Models (VLMs) [10,7,6], which have demonstrated strong perceptual and reasoning capabilities through large-scale multimodal pretraining, these models still struggle in real medical scenarios—frequently misinterpreting fine-grained pathological patterns, lacking necessary domain priors, and hallucinating confidently but incorrect diagnoses, posing unacceptable risks in safety-critical clinical applications [17].

[†]These authors contributed equally to this work.

A natural remedy is to fine-tune VLMs on medical instruction datasets [11]. However, this paradigm faces three intertwined obstacles. **First**, high-quality labeled medical data is scarce and expensive to curate, while full-parameter fine-tuning of large VLMs demands prohibitive computational resources. **Second**, due to these costs, many medical VLMs resort to substantially smaller backbones ($\leq$7B parameters), whose limited capacity constrains complex clinical reasoning even after domain adaptation [5]. **Third**, fine-tuning—even on larger models—can disrupt pretrained representations, increasing the risk of forgetting general competences and causing brittle generalization across medical sub-domains [15]. Recent efforts such as reinforcement learning for medical reasoning [18], multimodal contrastive learning [23,12], and thinking distillation for retrieval-augmented generation [8] achieve promising results, yet all fundamentally rely on gradient-based optimization and remain susceptible to the aforementioned limitations.

These challenges suggest that the bottleneck lies not in model architecture but in the reliance on parameter updates as the sole mechanism for acquiring medical expertise. This motivates a fundamental question: *Can we enhance the medical competency of generalist VLMs without any training, thereby avoiding both the computational burden and the risk of forgetting?* We address this question by introducing **EvoMed**, a training-free framework that enables a generalist VLM to acquire and apply medical knowledge through self-evolving learning. Inspired by recent advances in training-free policy optimization [3] and experience-driven agent evolution [4,21,19], EvoMed eschews parameter updates entirely; instead, it leverages in-context learning and retrieval-augmented generation to construct an external, dynamically evolving knowledge base ($\mathcal{KB}$) that accumulates distilled clinical insights over time. During a self-evolving phase, the VLM generates diverse reasoning trajectories, critiques its own errors, and distills concise medical priors stored as reusable knowledge entries. At inference time, EvoMed retrieves the most relevant priors and injects them into the VLM's context, guiding it toward clinically grounded, interpretable, and hallucination-resistant predictions.

Our contributions are summarized as follows:

- We propose **EvoMed**, a training-free medical VQA framework that enables generalist VLMs to acquire specialized clinical expertise through self-evolving knowledge accumulation, without modifying any model parameters.
- We introduce a coarse-to-fine retrieval strategy combining semantic triplet matching and vision-augmented cross-encoder reranking, bridging the modality gap between visual queries and textual knowledge entries.
- Extensive experiments on OmniMedVQA [9] and SLAKE [14] demonstrate that EvoMed, using only 500 evolution samples and approximately 15M tokens, consistently improves frozen same-backbone baselines, provides data-efficient performance compared with established medical VLMs, and exhibits strong cross-model knowledge transferability.

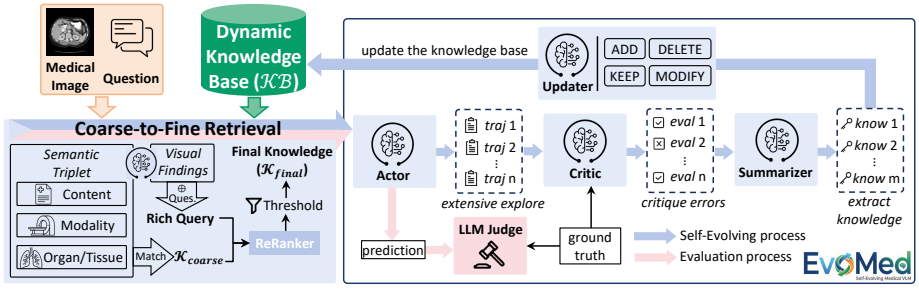

Fig. 1. The framework of EvoMed. EvoMed integrates two synergistic mechanisms: (1) *Retrieval-Augmented Strategy*, a coarse-to-fine strategy injecting relevant clinical priors; and (2) *Self-Evolving Knowledge Accumulation*, a closed-loop process where the VLM dynamically distills insights into the Knowledge Base ($\mathcal{KB}$).

2 Methodology

2.1 Overview

As illustrated in Fig. 1, EvoMed consists of two tightly coupled components: a **Coarse-to-Fine Retrieval** mechanism shared across both phases, and a **Self-Evolving Knowledge Accumulation** loop that populates the knowledge base.

2.2 Coarse-to-Fine Retrieval-Augmented Strategy

This mechanism systematically injects relevant clinical priors from the $\mathcal{KB}$ to ground the model's predictions. To bridge the vision-language modality gap, we employ a two-step strategy:

1. **Coarse Retrieval (Semantic Filtering):** The VLM jointly outputs a description of visual findings V_{desc} and a Semantic Triplet T (Query Content, Modality, Organ/Tissue). Query Content denotes the clinical intent of the question, such as presence, location, abnormality type, or diagnosis, rather than the answer itself. Because the triplet schema is discrete and low-dimensional, it provides a stable coarse filter that instantly prunes the $\mathcal{KB}$ to a topically relevant subset $\mathcal{K}_{coarse}$ by matching the stored triplets T_j.
2. **Fine-grained Reranking & Generation:** To resolve semantic ambiguities, a cross-encoder $\mathcal{R}(\cdot,\cdot)$ scores the alignment between each candidate knowledge $K_j \in \mathcal{K}_{coarse}$ and a concatenated query $[V_{desc} \oplus Q]$:

$$s_j = \mathcal{R}([V_{desc} \oplus Q], K_j) \tag{1}$$

Knowledge entries with scores above a threshold τ form the final set $\mathcal{K}_{final}$. Conditioned on this context, the VLM generates the final prediction:

$$A_{\text{final}} = \text{VLM}(I, Q, \mathcal{K}_{final}) \tag{2}$$

This explicit injection encourages clinically grounded outputs without parameter updates. The reranker also reduces the effect of residual errors in the open-ended V_{desc} by requiring candidate priors to align with both the visual description and the original question.

2.3 Self-Evolving Knowledge Accumulation

This mechanism acts as our training-free adaptation process, mimicking the human learning cycle of practice, reflection, and revision. Over a small adaptation set, the VLM operates in a closed loop with four distinct roles:

- **Actor (Reasoning):** For each sample (I, Q), the VLM retrieves relevant priors $\mathcal{K}_{final}$ from $\mathcal{KB}$ via the coarse-to-fine strategy (Sec. 2.2), then generates N diverse reasoning trajectories through stochastic sampling:

$$\{(R_i, A_i)\}_{i=1}^{N} \sim P_{\text{VLM}}(R, A \mid I, Q, \mathcal{K}_{final}) \tag{3}$$

- **Critic (Analysis):** Each trajectory is evaluated against the ground truth A_{gt}, and identifies why reasoning succeeded or failed:

$$C_i = \text{Critic}(Q, R_i, A_i, A_{gt}), \quad C_i = (\text{eval}_i, \text{analysis}_i) \tag{4}$$

- **Summarizer (Reflection):** Activated only when trajectories contain both correct and incorrect answers, the Summarizer distills a generalizable clinical prior K_{new} via contrastive comparison:

$$K_{new} = \text{Summarizer}\big(Q, \{(R_i, C_i)\}_{i=1}^{N}\big) \tag{5}$$

- **Updater (Maintenance):** Rather than blindly appending K_{new}, the Updater compares it against existing entries and selects one of four operations to keep $\mathcal{KB}$ compact and non-redundant:

$$\text{op} = \text{Updater}(K_{new}, \mathcal{KB}) \in \{\textsc{Add}, \textsc{Modify}, \textsc{Delete}, \textsc{Keep}\} \tag{6}$$

ensuring the knowledge base remains compact, accurate, and non-redundant.

2.4 Mechanism of Evolution and Capability Preservation

By decoupling knowledge acquisition from parameter optimization, EvoMed maintains strictly frozen model weights throughout its lifecycle, avoiding the parameter-update-induced forgetting risk that can arise in fine-tuning approaches and preserving the VLM's original generalist capabilities by construction.

Meanwhile, the closed-loop accumulation mechanism systematically externalizes the VLM's trial-and-error experience into refined clinical rules within $\mathcal{KB}$. During inference, the retrieval mechanism acts as a dynamic working memory, explicitly injecting domain-specific priors into the context window. This forces the model to ground its generation in retrieved clinical evidence rather than relying on implicit parametric memory, inherently bridging the domain gap and mitigating confident hallucinations.

3 Experiments

3.1 Datasets and Experimental Setup

To evaluate the effectiveness of EvoMed in a realistic, low-resource medical adaptation scenario, we conducted experiments on two prominent medical VQA benchmarks under the closed-set setting, which enables standardized and reproducible evaluation. We strictly limited the "evolution" (training-free learning) phase to a small subset of data to demonstrate data efficiency.

- **OmniMedVQA** [9]: A comprehensive medical VQA benchmark covering multiple modalities. We focus on the Radiology (Rad) and Microscopy (Mic) subsets.
  - *OmniMed-Rad:* Includes CT, X-ray, and MRI modalities. We randomly sampled 500 VQA pairs for the self-evolving knowledge accumulation phase and evaluated on a disjoint test set of 550 samples. CT, X-ray, and MRI are modality-stratified subsets of this test set rather than sequential tasks; Table 1 reports final accuracies after one $\mathcal{KB}$ is built from Rad-500.
  - *OmniMed-Mic:* Focuses on microscopy images. Similarly, we sampled 500 pairs (*Mic-500*) for evolution and evaluated on 534 test samples.
- **SLAKE** [14]: A bilingual dataset rich in semantic labels. We utilized a random subset of 500 QA pairs for evolution and evaluated performance on the official test split (416 samples).

OmniMed-Rad, OmniMed-Mic, and SLAKE are treated as independent adaptation settings, each with its own 500-sample evolution set and separately constructed $\mathcal{KB}$.

3.2 Implementation Details

We employed **Qwen2.5-VL-7B-Instruct** [2] and **Qwen3-VL-235B-A22B** [1] as our backbone VLMs, with `bge-reranker-v2-m3` [22] for fine-grained retrieval re-ranking and **Qwen3-Max** [20] as the LLM-as-Judge evaluator.

During the *Self-Evolving* phase, the Actor generated diverse reasoning paths (temperature 0.9), while the Critic used greedy decoding (temperature 0.0) for deterministic evaluation. The Summarizer and Updater operated at temperature 0.3. During *Inference*, the knowledge relevance threshold was set to $\tau = 0.8$, and the final answer was generated with temperature 0.0 for reproducibility.

3.3 Results and Analysis

Consistent Improvements over Baselines. As shown in Table 1, EvoMed improves the performance of both Qwen2.5-VL-7B and Qwen3-VL-235B-A22B across radiological modalities. The improvement is more pronounced for the 7B backbone, which achieves an average gain of +6.6%, including a +8.5% gain on CT. This suggests that smaller models, which may exhibit larger domain

Table 1. Performance comparison on OmniMedVQA-Rad across different imaging modalities. We compare frozen backbone models against their EvoMed-enhanced counterparts. CT, X-ray, and MRI are modality-stratified subsets of the same disjoint test set. The "Cost" column reports the total number of tokens (input + output) consumed during the self-evolving phase. The "Weighted Avg." is computed by weighting each modality proportionally to its number of test samples.

Method	Cost	Train Set	CT	X-Ray	MRI	Weighted Avg.
Qwen2.5-VL-7B	-	-	70.59	74.68	63.84	67.21
EvoMed (7B)	≈12M	Rad-500	**79.08**$_{+8.49}$	**79.75**$_{+5.07}$	**69.81**$_{+5.97}$	**73.82**$_{+6.61}$
Qwen3-VL-235B-A22B	-	-	80.39	84.81	78.93	80.18
EvoMed (A22B)	≈15M	Rad-500	**87.58**$_{+7.19}$	**88.61**$_{+3.80}$	**83.96**$_{+5.03}$	**85.64**$_{+5.46}$

gaps in specialized medical scenarios, can particularly benefit from the explicit incorporation of retrieved clinical priors.

Data Efficiency vs. Extensive Fine-Tuning. A key advantage of EvoMed is its ability to achieve highly competitive performance using orders of magnitude less adaptation data (Table 2). Comparisons with LLaVA-Med and Med-Flamingo should be interpreted as reference points rather than same-backbone comparisons, since these SFT medical VLMs use earlier backbone architectures and different training distributions. The controlled evidence for EvoMed is therefore the same-backbone improvement: with only 500 evolution samples, it raises the frozen Qwen2.5-VL-7B model from 67.21% to 73.82% and the Qwen3-VL-235B-A22B backbone from 80.18% to 85.64% on OmniMed-Rad.

Table 2. Performance comparison with representative medical VLMs on OmniMed-VQA and SLAKE datasets. "Num. of Seen Medical Samples" indicates the data volume used for domain adaptation (fine-tuning or evolution). Results from SFT medical VLMs are not same-backbone comparisons and should be interpreted with their different backbones and training distributions in mind.

Method	Num. of Seen Medical Samples	Accuracy (%)			
		OmniMed-Rad	OmniMed-Mic	SLAKE	Avg.
Zero-shot Baselines					
Qwen2.5-VL-7B	/	67.21	61.05	63.22	63.83
Qwen3-VL-235B-A22B	/	80.18	78.65	74.04	77.62
Zero-shot Medical VLM					
Huatuo-GPT-vision-34B [6]	≈1.3M	73.64	68.54	76.90	73.03
Supervised Fine-tuning (SFT)					
Med-Flamingo [16]	>100K	31.82	32.02	84.80	49.55
LLaVA-Med [11]	600K	52.91	49.63	**85.34**	62.63
Ours (Training-free)					
EvoMed (Qwen2.5-VL-7B)	**500**	73.82	72.47	67.79	71.36
EvoMed (Qwen3-VL-235B-A22B)	**500**	**85.64**	**85.21**	80.53	**83.79**

Reducing Over-specialization Risk. Extensive supervised fine-tuning (SFT) on narrow medical datasets can lead to brittle cross-dataset generalization. As demonstrated in Table 2, LLaVA-Med achieves high accuracy on SLAKE

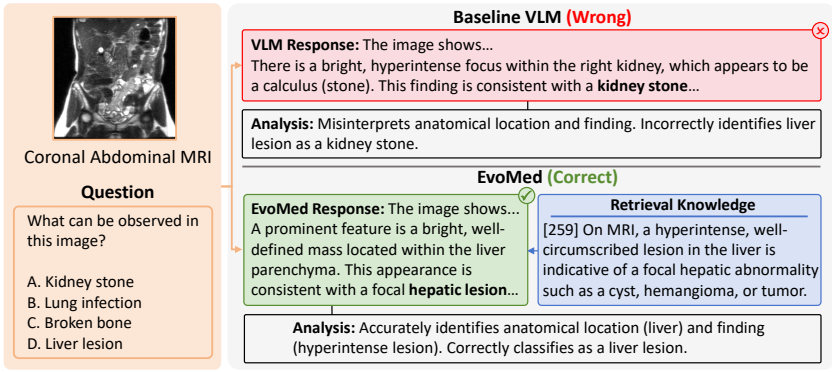

Fig. 2. Qualitative Case Study. Comparison of generated reasoning trajectories between the Baseline model and EvoMed on a coronal abdominal MRI.

(85.34%), which is closer to its training distribution, but performs much lower on OmniMed-Rad (52.91%). By keeping model weights strictly frozen and relying on self-evolving retrieval, EvoMed avoids parameter-update-induced forgetting risk by construction while improving its own frozen baselines across the evaluated datasets.

Table 3. Cross-model knowledge transfer performance. We evaluate how the Qwem2.5-VL-7B and Qwen3-VL-235B-A22B backbones perform when utilizing knowledge bases ($\mathcal{KB}$) generated by themselves versus those generated by the other model.

Inference Model	Knowledge Source	Accuracy (%)			
		OmniMed-Rad	OmniMed-Mic	SLAKE	Avg.
Qwen2.5-VL-7B	-	67.21	61.05	63.22	63.83
	$\mathcal{KB}_{7B}$	$73.82_{+6.61}$	$72.47_{+11.42}$	$67.79_{+4.57}$	$71.36_{+7.53}$
	$\mathcal{KB}_{A22B}$	$71.82_{+4.61}$	$72.28_{+11.23}$	$69.95_{+6.73}$	$71.35_{+7.52}$
Qwen3-VL-235B-A22B	-	80.18	78.65	74.04	77.62
	$\mathcal{KB}_{A22B}$	$85.64_{+5.46}$	$85.21_{+6.56}$	$80.53_{+6.49}$	$83.79_{+6.17}$
	$\mathcal{KB}_{7B}$	$83.09_{+2.91}$	$80.90_{+2.25}$	$75.72_{+1.68}$	$79.90_{+2.28}$

Cost-Efficiency for Clinical Deployment. Unlike traditional fine-tuning paradigms that require dedicated optimization runs, EvoMed's knowledge accumulation phase is highly cost-effective. As detailed in Table 1, evolving the Qwen2.5-VL-7B model on 500 samples for 3 epochs costs approximately $10 via API, offering an accessible and scalable solution for real-world clinical deployment.

Qualitative Case Study Fig. 2 presents a comparison on a coronal abdominal MRI case. The baseline model hallucinates a kidney stone by misinterpreting a hyperintense hepatic focus. In contrast, EvoMed retrieves relevant clinical priors

(e.g., *"a hyperintense, well-circumscribed hepatic lesion suggests a focal abnormality such as a cyst or hemangioma"*), correctly identifies the liver lesion, and produces a transparent, evidence-based reasoning chain, demonstrating the interpretability and reliability required for clinical scenarios.

3.4 In-depth Analysis and Ablation Studies

Evolution Dynamics and Scaling. We track test accuracy and knowledge base size across 3 evolution epochs for the Qwen3-VL-235B-A22B backbone (Fig. 3). EvoMed exhibits strong data efficiency: accuracy surges during the first epoch (e.g., SLAKE jumps from 74.04% to over 79% with merely 139 knowledge items) and plateaus by Epoch 3, while the knowledge base remains highly compact (e.g., 524 items for OmniMed-Rad). This demonstrates that the Critic-Summarizer pipeline swiftly captures core domain knowledge while effectively filtering redundancy.

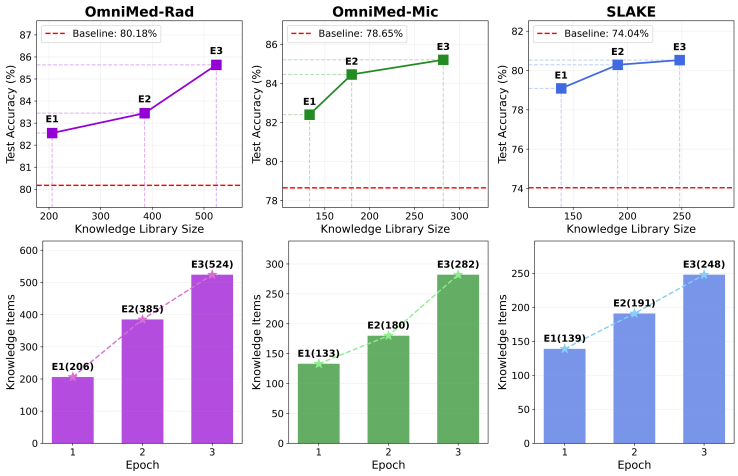

Fig. 3. Evolution Dynamics of EvoMed (Qwen3-VL-235B-A22B). The top row illustrates the correlation between Knowledge Library Size and Test Accuracy across 3 evolution epochs on OmniMed-Rad, OmniMed-Mic, and SLAKE. The bottom row displays the corresponding growth of accumulated knowledge items. EvoMed achieves significant performance gains rapidly with a highly compact knowledge base.

Cross-Model Knowledge Transfer. To investigate the universality of the accumulated clinical priors, we evaluate each model using the $\mathcal{KB}$ generated by its counterpart (Table 3). We observe an asymmetric but consistently positive transferability. For the Qwen2.5-VL-7B model, leveraging $\mathcal{KB}_{\text{A22B}}$ yields comparable average gains (+7.52%) to using its own $\mathcal{KB}_{\text{7B}}$ (+7.53%), with notably stronger improvement on SLAKE (69.95% vs. 67.79%), suggesting that the larger Qwen3-VL-235B-A22B backbone distills more precise and generalizable clinical

rules. Conversely, while $\mathcal{KB}_{7B}$ still provides consistent gains for the Qwen3-VL-235B-A22B backbone (+2.28% on average), the improvement is less pronounced than using $\mathcal{KB}_{A22B}$ (+6.17%). This confirms that the distilled knowledge base acts as a model-agnostic semantic prior: even a weaker model's accumulated experience can benefit a stronger one, while stronger models naturally produce higher-quality priors that more effectively bridge reasoning capability gaps across scales.

4 Conclusion

We presented EvoMed, a training-free framework that adapts generalist VLMs to the medical domain without any parameter updates. By leveraging a self-evolving mechanism, EvoMed iteratively distills and accumulates clinical priors into a dynamic knowledge base using minimal data. During inference, a coarse-to-fine retrieval strategy explicitly injects this knowledge to guide accurate and interpretable clinical reasoning. Experiments demonstrate consistent same-backbone gains, strong data efficiency, and cross-model knowledge transfer, while the frozen-backbone design avoids parameter-update-induced forgetting risk. Future work will explore formal continual-learning forgetting metrics and extension to complex 3D medical imaging modalities.

Acknowledgments. This work was partially supported by the AI4S Initiative and HPC Platform of ShanghaiTech University.

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

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. 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. arXiv preprint arXiv:2502.13923 (2025)
3. Cai, Y., Cai, S., Shi, Y., Xu, Z., Chen, L., Qin, Y., Tan, X., Li, G., Li, Z., Lin, H., et al.: Training-free group relative policy optimization. arXiv preprint arXiv:2510.08191 (2025)
4. Cai, Z., Guo, X., Pei, Y., Feng, J., Su, J., Chen, J., Zhang, Y.Q., Ma, W.Y., Wang, M., Zhou, H.: Flex: Continuous agent evolution via forward learning from experience. arXiv preprint arXiv:2511.06449 (2025)
5. Chen, J., Jiang, Y., Yang, D., Li, M., Wei, J., Qian, Z., Zhang, L.: Can llms' tuning methods work in medical multimodal domain? In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 112–122. Springer (2024)

6. Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Cai, Z., Ji, K., Wan, X., et al.: Towards injecting medical visual knowledge into multimodal llms at scale. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 7346–7370 (2024)
7. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
8. Dai, P., Ou, Y., Yang, Y., Jin, Z., Suzuki, K.: Goca: Trustworthy multi-modal rag with explicit thinking distillation for reliable decision-making in med-lvlms. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 251–261. Springer (2025)
9. 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)
10. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
11. 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)
12. Liang, X., Wang, Y., Wang, D., Jiao, Z., Zhong, H., Yang, M., Wang, Q.: Leveraging coarse-to-fine grained representations in contrastive learning for differential medical visual question answering. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 415–425. Springer (2024)
13. Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., Ge, Z.: Medical visual question answering: A survey. Artificial Intelligence in Medicine **143**, 102611 (2023)
14. 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)
15. Luo, Y., Yang, Z., Meng, F., Li, Y., Zhou, J., Zhang, Y.: An empirical study of catastrophic forgetting in large language models during continual fine-tuning. IEEE Transactions on Audio, Speech and Language Processing (2025)
16. 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)
17. Nam, Y., Kim, D.Y., Kyung, S., Seo, J., Song, J.M., Kwon, J., Kim, J., Jo, W., Park, H., Sung, J., et al.: Multimodal large language models in medical imaging: current state and future directions. Korean Journal of Radiology **26**(10), 900 (2025)
18. 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)

19. Sun, Z., Liu, Z., Zang, Y., Cao, Y., Dong, X., Wu, T., Lin, D., Wang, J.: Seagent: Self-evolving computer use agent with autonomous learning from experience. arXiv preprint arXiv:2508.04700 (2025)
20. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
21. Yang, C., Yang, X., Wen, L., Fu, D., Mei, J., Wu, R., Cai, P., Shen, Y., Deng, N., Shi, B., et al.: Learning on the job: An experience-driven self-evolving agent for long-horizon tasks. arXiv preprint arXiv:2510.08002 (2025)
22. Yang, T.L., Liu, J.S., Tseng, Y.H., Jang, J.S.R.: Knowledge retrieval based on generative ai. arXiv preprint arXiv:2501.04635 (2025)
23. Yazici, Z.A., Ekenel, H.K.: Bamco: Balanced multimodal contrastive learning for knowledge-driven medical vqa. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 77–87. Springer (2025)