MedExpMem: Adapting Experience Memory for Differential Diagnosis

Qianhan Feng[1*], Zhongzhen Huang[2*], Yakun Zhu[2,3], Yannian Gu[2], Winnie Chiu Wing CHU[1], Xiaofan Zhang[2,3✉], and Qi Dou[1✉]

[1] The Chinese University of Hong Kong
[2] Shanghai Jiao Tong University
[3] Shanghai Innovation Institue
qianhan.feng@link.cuhk.edu.hk qidou@cuhk.edu.hk
*Equal contribution. ✉Corresponding authors.

Abstract. Experienced physicians develop diagnostic expertise through clinical practice, acquiring not only disease knowledge but also the ability to differentiate confusable conditions. Current medical vision-language models (VLMs) lack this capability—their parameters encode static knowledge that does not evolve across diagnostic encounters. We propose **MedExpMem**, an experience memory framework enabling VLM-based diagnostic agents to accumulate differential diagnosis expertise. Unlike retrieval-augmented generation, which retrieves encyclopedic disease descriptions, MedExpMem distills discriminative experience derived from the agent's own diagnostic failures with guidance and organizes them as pairwise differential notes encoding key discriminators, actionable decision rules and reasoning error patterns. The framework adopts a two-phase construction process mirroring physician learning: initial practice exposes knowledge gaps, and reflective re-diagnosis refines understanding. When encountering new cases, the agent retrieves experience memory to guide differential reasoning. We evaluate MedExpMem on a radiology benchmark spanning 11 subspecialties. Results demonstrate consistent accuracy improvements, maximum 7.0%, across diverse models and scales. Analytical experiments validate experience quality and robustness, demonstrating that structured experience addresses adaptation needs beyond the reach of parametric learning alone.

Keywords: Agent Memory · Multi-modality · Differential Diagnosis

1 Introduction

Medical expertise is not acquired solely through textbook knowledge; it is refined through repeated clinical practice. Although trainees may master canonical disease descriptions—typical imaging patterns, diagnostic criteria, and management guidelines—they often struggle to match the performance of experienced clinicians. The gap lies in clinical experience: discriminative, case-derived insights accumulated through repeated exposure to diagnostic ambiguity [1]. Unlike textbook knowledge, which describes diseases in isolation, clinical experience encodes

comparative judgments—why two conditions are easily confused, which subtle features distinguish them, and when one diagnosis should be favored. Such experience is not a verbatim archive of past cases but an abstraction distilled from prior diagnostic successes and errors.

Vision-language models (VLMs) have demonstrated strong capabilities in medical image interpretation [2, 3], yet operate in a fundamentally stateless manner: each diagnostic session proceeds independently without access to prior encounters. Models do not accumulate persistent experience from ongoing deployment. While fine-tuning could in principle encode experiential knowledge into parameters, it remains impractical in clinical settings due to privacy constraints, computational costs, and catastrophic forgetting risks—particularly for locally deployed hospital models requiring continuous adaptation.

Existing approaches address related aspects but do not enable persistent accumulation of differential experience. Retrieval-augmented generation (RAG) supplements models with external knowledge [4], yet standard retrieval favors encyclopedic disease descriptions over the discriminative knowledge essential for differential diagnosis. Memory mechanisms for LLM-based agents [5, 6, 12, 11, 7] typically store conversation histories or factual summaries rather than abstracting comparative reasoning patterns. Self-correction frameworks including Self-RAG [14] and Reflexion [10] enable within-session improvement but lack persistent, cross-case memory. Neither paradigm addresses what medical diagnosis specifically demands: structured pairwise experience encoding why conditions are confused and how to distinguish them.

To bridge this gap, we propose **MedExpMem**, an experience memory framework enabling VLM-based diagnostic agents to accumulate differential expertise without parameter updates. Rather than organizing memory around isolated diseases, MedExpMem encodes knowledge at the level of diagnosis pairs—the fundamental unit of differential reasoning. Each pairwise experience note is derived from the agent's own diagnostic failures, capturing expert-guided experience tailored to its specific blind spots. Notes include **key discriminators**, **actionable decision rules**, and **reasoning error patterns** that led to misdiagnosis. Critically, MedExpMem differs from RAG not only in what it retrieves, but in how memory is constructed: extracted from self-failures rather than external corpora.

Experience memory construction comprises two phases, both guided by expert-verified diagnostic knowledge. In **Phase I**, the agent performs zero-shot diagnosis to expose its intrinsic reasoning blind spots. In **Phase II**, the agent revisits erroneous cases with experience accumulated from Phase I, simulating iterative practice to consolidate reliable patterns, correct inconsistent reasoning, and filter spurious errors. At inference, relevant experience is retrieved to help the agent recognize differential pitfalls it tends to overlook. This process mirrors clinical learning: initial exposure reveals weaknesses, deliberate practice refines judgment, and accumulated experience persists across cases. By decoupling experience from parameters, MedExpMem enables continual adaptation in privacy-sensitive medical environments.

We construct a benchmark from Eurorad [17], an educational radiology repository providing case data and expert guidance. This education-oriented setting closely mirrors experience accumulation scenarios and poses greater diagnostic complexity. Experiments across multiple VLM families including Qwen3-VL demonstrate consistent improvements. Our contributions are threefold:

- A structured pairwise experience schema capturing discriminators, decision rules and reasoning error patterns—organized around diagnosis pairs to directly support differential reasoning.
- A two-phase memory construction process that extracts and refines experience from diagnostic errors, enabling continuous and robust accumulation without parameter updates.
- Comprehensive evaluation on an education-oriented radiology benchmark demonstrating consistent improvements, with analytical experiments validating experience quality and robustness.

2 MedExpMem Framework

2.1 Problem Formulation

Given a patient's medical images I and clinical history x, a VLM-based diagnostic agent $\mathcal{A}$ renders a diagnosis $d = \mathcal{A}(x, I)$. This formulation is static, where agent knowledge remains unchanged regardless of cases encountered. We extend this to a dynamic adapting paradigm:

$$d = \mathcal{A}(x, I; \mathcal{M}), \quad e = \textsc{Extract}(x, I, d, d^*), \quad \mathcal{M} \leftarrow \mathcal{M} \cup \{e\} \tag{1}$$

where the agent first produces diagnosis d conditioned on experience memory $\mathcal{M}$, then compares d against expert ground truth d^* to extract and accumulate experience memory e. This formulation enables continuous experience accumulation without parameter updates—particularly valuable for locally deployed models where fine-tuning is impractical due to privacy and computational constraints.

2.2 Pairwise Differential Experience Notes

We design structured experience notes organized around *diagnosis pairs*, the fundamental unit of differential reasoning, rather than isolated diseases. This pairwise organization offers critical advantages over single-disease retrieval. First, it directly addresses differential diagnosis: when distinguishing confusable conditions, an "A v.s. B" note provides targeted discriminative knowledge unavailable from separate descriptions of A and B. Second, it reflects the clinical reasoning pattern "Is this A or B" rather than "What is A" [1]. Third, it supports bidirectional reasoning: each note encodes features favoring *both* diagnoses, preventing confirmation bias that single-disease retrieval may introduce.

Table 1 summarizes the note schema for experience memory. Each note is indexed by `differential_pair`, e.g., "Lymphoma v.s. Metastasis", representing the minimal discriminative unit in differential diagnosis. Core fields include

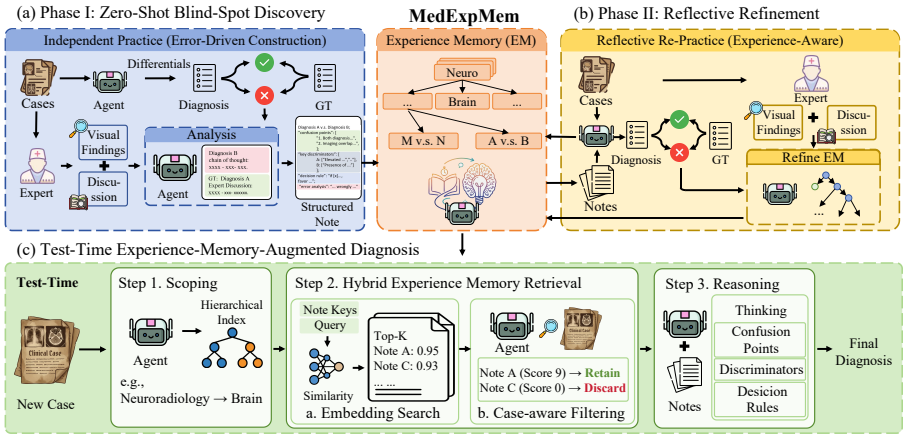

Fig. 1. Overview of MedExpMem. **(a) Phase I: Zero-Shot Blind-Spot Discovery.** The agent conducts zero-shot diagnosis. **(b) Phase II: Reflective Refinement.** The agent re-diagnoses cases with experience memory access. **(c) Test-Time Inference.** Agent performs experience-memory-augmented reasoning with hybrid-retrieval.

`discriminators` capturing distinguishing features, `decision_rule` providing actionable conditional guidance, and `error_analysis` recording agent's own reasoning error patterns that led to past misdiagnosis. Unlike static knowledge bases describing diseases in isolation, these fields are derived from agent-specific failures, yielding experience tailored to the model's own blind spots. Decision rules adopt probabilistic formulations to acknowledge diagnostic uncertainty, and are directly applicable as conditional checks during agent reasoning.

The experience memory notes are organized following standard radiology subspecialty divisions [25]: the first level corresponds to **departments** (e.g., neuroradiology, thoracic), and the second level specifies **organs/regions** within each department. This mirrors clinical radiology structure and enables anatomically-scoped retrieval. Our taxonomy covers 11 departments and 118 organ/region categories, with `others` providing coverage for atypical cases.

2.3 Two-Phase Experience Memory Construction

We construct experience memory through a two-phase process: initial practice exposing knowledge gaps, followed by reflective refinement.

Phase I: Zero-shot Error Discovery. The agent first diagnoses educational cases without memory access, exposing intrinsic reasoning limitations. For each error where diagnosis $d \neq d^*$, we extract an experience note:

$$e = \textsc{Extract}(x, I, d, d^*, \text{discussion}) \tag{2}$$

where `discussion` contains expert imaging findings and diagnostic reasoning. The agent is instructed to: (1) identify generalizable confusion points between d

Table 1. Experience note schema for differential diagnosis.

Field	Description [Format]
Metadata	
`department`	Radiology subspecialty for hierarchical indexing [single value]
`organ/region`	Anatomical localization within department [single value]
`differentials`	Two diagnoses being compared, serving as retrieval key ["A v.s. B"]
Core Differential Knowledge	
`confusions`	Reasons why these diagnoses are commonly confused—overlapping imaging features and clinical presentations [enumerated list]
`discriminators`	Imaging, clinical, and laboratory findings distinguishing each diagnosis [structured text with disease-specific sections]
`decision_rule`	Actionable conditional rules ["If $X_1 \rightarrow$ favor A; If $X_2 \rightarrow$ exclude A; If $Y_1 \rightarrow$ favor B; If $Y_2 \rightarrow$ exclude B; If Z $\rightarrow$ consider others"]
Self Reasoning Error Pattern	
`error_analysis`	Case-specific reasoning failure decomposition

and d^*, (2) extract discriminating features, (3) formulate bidirectional decision rules, and (4) analyze its reasoning failure. Memory inaccessibility during Phase I helps ensure notes reflect genuine knowledge gaps.

Phase II: Reflective Refinement. The second phase re-diagnoses all cases with memory access, serving complementary purposes: (1) *capturing stochastic errors*, cases correctly diagnosed by chance in may fail under repeated trials; (2) *identifying note deficiencies*, when relevant notes are retrieved but errors persist, Phase II extraction supplements missing information; (3) *detecting misleading notes*, if notes cause correct-to-incorrect transitions, error analysis enables refinement. New insights are merged with existing ones if differential pairs exist.

2.4 Experience-Memory-Augmented Diagnosis

When diagnosing new cases, the agent performs diagnosis through three steps:

Step 1: Anatomical Scoping. Clinical cases involve multiple organ systems or cross departmental boundaries. The agent recalls experience memory's hierarchical index and selects 1–2 department-organ paths based on case presentation.

Step 2: Hybrid Experience Retrieval. Within selected paths, embedding similarities between candidate differential diagnoses and each note's `differential pair` keys are computed using a domain-specific encoder. Pairwise matching ensures symmetric relevance to both competing hypotheses, mitigating retrieval bias toward a single diagnosis. Notes exceeding similarity threshold τ are retrieved, and top-K truncation is applied. The agent then scores each retrieved note on relevance, excluding clearly irrelevant notes (wrong organ system, unrelated disease category) while retaining potentially useful ones—a conservative strategy robust to smaller model limitations.

Step 3: Experience-Memory-Augmented Reasoning. Retained experience memories are integrated into the diagnostic prompt. The agent autonomously leverages notes to inform its differential reasoning and produce the final diagnosis.

3 Experiments

3.1 Experimental Setup

Dataset. We construct our benchmark from Eurorad [17], an open-source peer-reviewed radiology teaching file maintained by the European Society of Radiology. Eurorad comprises over 10,000 peer-reviewed cases spanning neuroradiology, thoracic, abdominal, musculoskeletal, cardiac, and interventional radiology. Each case provides structured *clinical history*, *multi-modality imaging* (CT, MRI, X-ray, ultrasound), *expert-annotated imaging findings*, *reference-based discussion*, *ground-truth diagnosis*, and *curated differential diagnoses* representing clinically plausible alternatives. Compared to classical medical VQA benchmarks [18–21], this education-oriented data with broader coverage is well-suited for targeted clinical experience accumulation. We collect 6,228 high-quality cases published before year 2025 for the *construction phase*, during which agents build personalized experience libraries. For evaluation, we curate 227 high-quality cases from year 2025 as test set. This temporal split mirrors realistic clinical training: accumulating experience from historical cases before encountering new challenges.

Implementation Details. We employ S-PubMedBert-MS-MARCO [15], a domain-specific embedding model built upon PubMedBERT [16], for keyword embedding. We set similarity threshold $\tau = 0.9$ and retrieve top-$K = 10$ notes. All VLMs operate with default temperature settings for 3 trials. During construction, each agent independently builds its own experience memory, ensuring notes align with agent-specific knowledge gaps.

3.2 Main Results

Medical deployment demands privacy protection and local inference, favoring open-source models. We evaluate Qwen3-VL [22], InternVL3.5 [23], and Lingshu [24], a medical-specialized VLM, across multiple scales.

Table 2 reports diagnostic accuracy with and without experience memory. MedExpMem consistently improves performance across all models, demonstrating that structured experience complements parametric knowledge regardless of architecture or scale. Even the strongest baseline gains +4.7%, showing experience addresses limitations beyond pre-training. Notably, the medical-specialized Lingshu-7B achieves lower baseline accuracy than subsequently released general-purpose models of comparable scale (Qwen3-VL-8B: 69.6%). This observation highlights a limitation of static domain-specific fine-tuning: models trained on historical medical data may underperform on challenging cases that emerge beyond their knowledge cutoff.

Table 2. Main results. **Baseline**: zero-shot diagnosis. **w/Exp**: diagnosis augmented with experience memory. **Recall**: proportion of cases with successfully retrieved notes. **Precision**: accuracy on cases where notes were retrieved. **Beneficial**: proportion of cases always corrected by experience memory (wrong → correct). **Harmful**: proportion of cases always misled by experience memory (correct → wrong).

Domain	Model	Diagnostic Accuracy			Experience Memory			
		Baseline	w/Exp	Δ	Recall	Precision	Beneficial	Harmful
General	Qwen3-VL-2B	54.6%	61.6%	**+7.0%**	49.6%	60.7%	8.4%	1.3%
	Qwen3-VL-8B	69.6%	74.5%	**+4.9%**	77.5%	74.4%	9.7%	4.8%
	Qwen3-VL-30B	74.9%	79.6%	**+4.7%**	58.6%	80.5%	7.0%	2.2%
	InternVL3.5-8B	64.0%	71.0%	**+6.9%**	81.9%	73.7%	12.7%	5.8%
	InternVL3.5-30B	74.0%	77.5%	**+3.5%**	68.3%	75.5%	8.8%	5.2%
Medical	Lingshu-7B	66.1%	67.8%	**+1.7%**	58.2%	68.2%	3.5%	1.7%
	Lingshu-32B	72.2%	74.8%	**+2.6%**	54.1%	74.2%	6.5%	3.8%

Table 3. Ablation study on Qwen3-VL-8B. Impact of construction rounds, retrieval scope, and similarity threshold.

Configuration	Accuracy	Δ
Baseline (zero-shot)	69.6%	–
w/ PubMed-RAG	72.4%	+2.8%
w/ **MedExpMem**	**74.5%**	**+4.9%**
MedExpMem Ablation		
One-round construction	73.1%	+3.5%
Single-department retrieval	72.2%	+2.6%
Threshold $\tau = 0.80$	73.1%	+3.5%
Threshold $\tau = 0.95$	72.8%	+3.2%

3.3 Analytical Experiments

We conduct ablation experiments on Qwen3-VL-8B to verify key design choices, and the results are reported in Table 3. ***Two-round construction*** further improves performance by expanding and refining notes. ***Cross-department search*** retrieval yields larger gains than single-department retrieval, reflecting realistic diagnostic complexity. For ***retrieval threshold***, performance peaks at $\tau = 0.9$, validating the similarity threshold design.

Furthermore, we compare the performance with RAG technology. Specifically, we implement retrieval from PubMed [26] and allow the agent to retrieve the top-3 related academic literature articles. However, the results show that textbook-like references provide limited gains to the agent when dealing with high challenging cases, as they increase retrieval scope and inference cost. This supports our hypothesis that agents benefit from customized experience memory to address their reasoning blind spots, rather than static textbook knowledge.

The experimental results in Table 2 verify the effectiveness of experience memory. We observe a recall–scale trade-off: larger models retrieve fewer notes

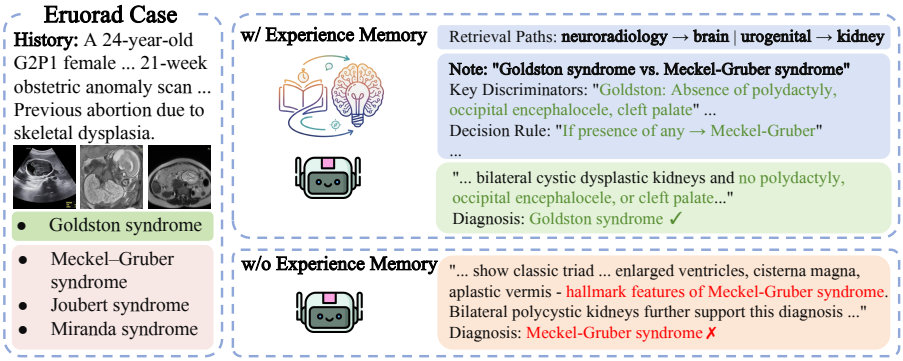

Fig. 2. Case study comparing diagnosis with and without experience memory. The retrieved pairwise note provides actionable discriminators that guide correct diagnosis.

due to fewer prior errors, whereas smaller models sometimes fail to identify optimal retrieval paths. Cases with retrieved notes are typically more challenging, yet experience memory elevates their accuracy toward baseline levels. Although memory can introduce both beneficial and harmful effects, the overall gains outweigh the risks, which largely depend on note quality and retrieval precision.

3.4 Case Study

Figure 2 shows a fetal anomaly case with Dandy–Walker malformation and cystic kidneys, where the key differential lies between Goldston syndrome and Meckel–Gruber syndrome. With experience memory, the agent retrieved the pairwise note "Goldston v.s. Meckel–Gruber", specifying that the absence of polydactyly, occipital encephalocele, and cleft palate favors Goldston syndrome. Recognizing these, the agent correctly diagnosed Goldston syndrome. Without memory, the agent focused on shared positive findings and predicted Meckel–Gruber syndrome. This illustrates how pairwise experience memory guides differential reasoning, which can be difficult to achieve utilizing finetuning.

4 Conclusion

We introduced MedExpMem, an experience memory framework that enables vision-language diagnostic agents to accumulate structured differential expertise without parameter updates. By organizing knowledge as pairwise differential notes rather than isolated disease descriptions, MedExpMem shifts retrieval from static characterization to comparative clinical reasoning. A two-phase construction process identifies reasoning blind spots and refines them through reflective re-diagnosis, allowing experience to persist across cases. Experiments on a temporally split radiology benchmark demonstrate consistent improvements across model families and scales, demonstrating that structured experience addresses adaptation needs beyond the reach of parametric knowledge alone.

Acknowledgement

This work was supported in part by the Hong Kong Innovation and Technology Fund (Project No. GHP/167/22SZ), and in part by a grant from the ANR/RGC Joint Research Scheme sponsored by the Research Grants Council of the Hong Kong Special Administrative Region, China and the French National Research Agency (Project No. A-CUHK402/23).

Disclosure of Interests

No competing interests.

References

1. Schmidt, H.G., Norman, G.R., Boshuizen, H.P.: A cognitive perspective on medical expertise: theory and implications. Academic Medicine **65**(10), 611–621 (1990)
2. Singhal, K., et al.: Large language models encode clinical knowledge. Nature **620**(7972), 171–180 (2023)
3. Zhang, Y., et al.: Large language models in medicine. Nature Medicine **29**(8), 1930–1940 (2023)
4. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. NeurIPS (2020)
5. Park, J.S., et al.: Generative agents: Interactive simulacra of human behavior. UIST (2023)
6. Zhong, W., et al.: MemoryBank: Enhancing large language models with long-term memory. AAAI (2024)
7. Zhao, A., et al.: ExpeL: LLM agents are experiential learners. AAAI (2024)
8. Mnih, V., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015)
9. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. NeurIPS (2017)
10. Shinn, N., et al.: Reflexion: Language agents with verbal reinforcement learning. NeurIPS (2023)
11. Zhang, Z., et al.: A survey on the memory mechanism of large language model based agents. arXiv preprint arXiv:2404.13501 (2024)
12. Maharana, A., et al.: Evaluating very long-term conversational memory of LLM agents. ACL (2024)
13. Xu, W., et al.: A-MEM: Agentic memory for LLMs. arXiv preprint arXiv:2502.12110 (2025)
14. Asai, A., et al.: Self-RAG: Learning to retrieve, generate, and critique through self-reflection. ICLR (2024)
15. Deka, P., et al.: S-PubMedBert-MS-MARCO: An efficient embedding model for biomedical information retrieval. arXiv preprint (2022)
16. Gu, Y., et al.: Domain-specific language model pretraining for biomedical natural language processing. ACM Trans. Comput. Healthcare **3**(1), 1–23 (2021)
17. Eurorad, https://www.eurorad.org, last accessed 2026/02/26

18. He, X., Zhang, Y., Mou, L., Xing, E., and Xie, P. PathVQA: 30000+ questions for medical visual question answering. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, pp. 485–495. Springer, 2020.
19. Lau, J. J., Gayen, S., Ben Abacha, A., and Demner-Fushman, D. A dataset of clinically generated visual questions and answers about radiology images. Scientific Data, 5(1):1–10, 2018.
20. Liu, B., et al.: SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. ISBI (2021)
21. Zhang, X., et al.: PMC-VQA: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415 (2023)
22. Bai, S., Cai, Y., Chen, R., et al.: Qwen3-VL Technical Report. arXiv preprint arXiv:2511.21631 (2025)
23. Wang, W., et al.: InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. arXiv preprint arXiv:2508.18265 (2025)
24. Xu, W., et al.: Lingshu: A Generalist Foundation Model for Unified Multimodal Medical Understanding and Reasoning. arXiv preprint arXiv:2506.07044 (2025)
25. American Board of Radiology: Subspecialty Certifications in Diagnostic Radiology. https://www.theabr.org/get-certified/subspecialties/, last accessed 2026/02/26
26. National Library of Medicine: PubMed. https://pubmed.ncbi.nlm.nih.gov/, last accessed 2026/02/26