

V-RAM: A Visual Retrieval-Augmented Multi-Agent Reasoning Framework for Medical Dialogue Diagnosis

Shukraditya Bose¹ Soumen Sinha^{2*} Ananya Bhat¹ Rahul Roy³

¹Vellore Institute of Technology, Vellore, India

²Delft University of Technology, Delft, The Netherlands

³Mahindra University, Hyderabad, India

*s.sinha-6@student.tudelft.nl

Abstract

Large Vision-Language Models (LVLMs) are increasingly explored for multimodal medical diagnosis, their ability to conduct accurate consultation remains challenging because of unstable visual reasoning and limited tracking of patient information across dialogue turns. To address this problem we present **V-RAM**(Visual Retrieval-Augmented Multi-Agent Reasoning Framework), a training-free framework for improving LVLM-based multimodal medical dialogue diagnosis. V-RAM improves medical dialogue diagnosis by adding visual retrieval, structured memory, and multi-agent reasoning around a pre-trained LVLM. It retrieves similar medical images as diagnostic cues and tracks patient evidence across dialogue turns. The framework operates at inference time, it requires no supervised fine-tuning and can be used with different LVLM backbones. Experiments on benchmark show that V-RAM outperforms existing baselines and the current LVLMs by **7.3** macro-F1 points on average with **76%** reduction in number of utterances on average that are used for medical dialogue diagnosis.

1 Introduction

Large Vision-Language Models (LVLMs) have shown promise for telemedicine consultation that require reasoning over both clinical images and patient text. It allows them to inspect visual findings, interpret complaints, and generate clinically relevant responses. However, medical consultation is not a single-turn question-answering task. A reliable consultation system must ask follow-up questions, track symptoms across turns, reason under incomplete information, and avoid premature or unsupported conclusions. This makes multimodal medical dialogue diagnosis more challenging than standard visual question answering or single-step disease classification.

A key challenge in this setting is that LVLMs can suffer from unstable visual reasoning and diagnostic drift across turns. In a multi-turn conversation, the model may over-rely on recent patient responses while forgetting earlier visual or textual evidence. It may also hallucinate diagnoses that are not visually supported by the medical image. These problems are especially important in telemedicine-like settings, where everyday clinical images may be noisy, patient descriptions may be incomplete, and the system must decide when enough information has been collected to commit to a diagnosis. In this work, we ask whether diagnostic reasoning can be made more stable by grounding the model’s decisions in visually plausible disease candidates before the dialogue begins, and by preserving clinically relevant information in a structured form as the consultation progresses.

The main objective of this paper is therefore to investigate how medical dialogue diagnosis can be improved without fine-tuning the underlying LVLM. We aim to solve three connected issues: first, how to provide the model with stronger visual diagnostic priors; second, how to prevent the loss of important patient information across dialogue turns; and third, how to reduce unsupported final diagnoses that are not grounded in the image or collected symptoms. By focusing on these issues, we propose **V-RAM**: A retrieval-augmented multi-agent reasoning framework for medical dialogue diagnosis.

2 Related Work

Most medical NLP benchmarks evaluate factual knowledge through static Question Answering tasks, MedQA (Jin et al., 2020) tests licensing exam questions, MedMCQA (Pal et al., 2022) covers 194K multiple-choice items, and Pub-MedQA (Jin et al., 2019) probes yes/no reasoning

over biomedical abstracts. While useful, none of these capture the interactive, information-seeking nature of a real consultation. MediQ (Li et al., 2024a) took a step in this direction by requiring models to ask clarifying questions before committing to a diagnosis. Sviridov et al. (Sviridov et al., 2025) extends by incorporating both image and text input into medical consultation.

With advancements in generative models the use of RAG (Lewis et al., 2020) becomes eminent as with a non-parametric retrieval index substantially improves performance on knowledge-intensive tasks. Subsequent work explored dense passage retrieval, fusing multiple retrieved passages (Izacard and Grave, 2021), and domain-specific retrieval for clinical QA (Xiong et al., 2024). RAG based systems use text without visual priors making the consultation process incomplete.

Recent work has explored decomposing medical reasoning across multiple collaborating agents. MDAgents (Kim et al., 2024) routes queries to different agent configurations based on task complexity, AgentClinic (Schmidgall et al., 2024) introduces role-playing doctor-patient-nurse agents in a simulated clinical environment, and AI Hospital (Fan et al., 2024) benchmarks LLMs across examination, diagnosis, and treatment. None of these systems maintain persistent structured memory across turns or constrain diagnosis to a retrieval-initialized candidate set.

Taken together, prior work has treated medical QA, retrieval augmentation, multi-agent coordination, and vision-language modeling as largely separate problems. To address the gap, we propose a V-RAM which is the first system (to the best of the knowledge of authors) that integrates all four into a single inference-time pipeline for multimodal medical dialogue.

3 Methodology

In this article, we propose **V-RAM** (Visual Retrieval-Augmented Multi-Agent Reasoning Framework), a training-free framework to improve LVLMM-based multimodal medical dialogue diagnosis. V-RAM enhances a pre-trained LVLMM with visual retrieval, structured memory, and multi-agent reasoning for more robust diagnostic reasoning. The framework retrieves clinically similar medical images as auxiliary diagnostic evidence and keeps patient-specific contextual

memory over multiple dialogue turns. The proposed model works at inference time, no supervised fine-tuning is required, and can be easily plugged into different LVLMM backbones. The block diagram of the proposed model is shown in Figure 1. Each of the component of the model is described in following subsections.

3.1 Task Definition

The LVLMMs are evaluated on multi-turn medical diagnosis through interactive doctor patient dialogue. Given a medical image P_k , an initial basic complaint BC_k , and a set of disease labels $\mathcal{L}$, the system engages in at most T dialog turns with a simulated patient agent to produce a final diagnosis $\hat{D}_k \in \mathcal{L}$. Each case provides:

- A ground-truth diagnosis $y \in \mathcal{L}$
- A medical image P_k (dermoscopy or clinical photograph)
- A basic complaint C_0 (initial symptom description)
- Hidden additional complaints H (revealed progressively through dialogue)
- Patient temperament type $\phi \in \{\text{sanguine, choleric, melancholic, phlegmatic}\}$

At each dialogue turn $t \in \{1, \dots, T\}$, the Doctor Agent receives the patient’s response to the previous query and must decide whether to (a) ask follow up questions to elicit additional symptoms from AC_k , or (b) commit to a diagnosis $\hat{D}_k$ and provide treatment recommendations.

3.2 Multi-Agent Architecture

The diagnostic process is decomposed into a framework of specialised agents coordinated by a central controller. All agents are implemented as frozen prompt-based modules over a shared LVLMM backbone. No model weights are fine-tuned. Given an input case (P_k, C_0, y) , the pipeline proceeds in three phases.

Initialisation: The Visual Prior Agent analyses (P_k, C_0) and proposes top-3 candidate diagnoses. Concurrently, the CLIP Retrieval Index is queried to retrieve the top-3 most frequent and visually similar diagnostic cues. Both sets of labels are stored in a Structured Memory and unified into Candidate Label Set $\mathcal{C}$.

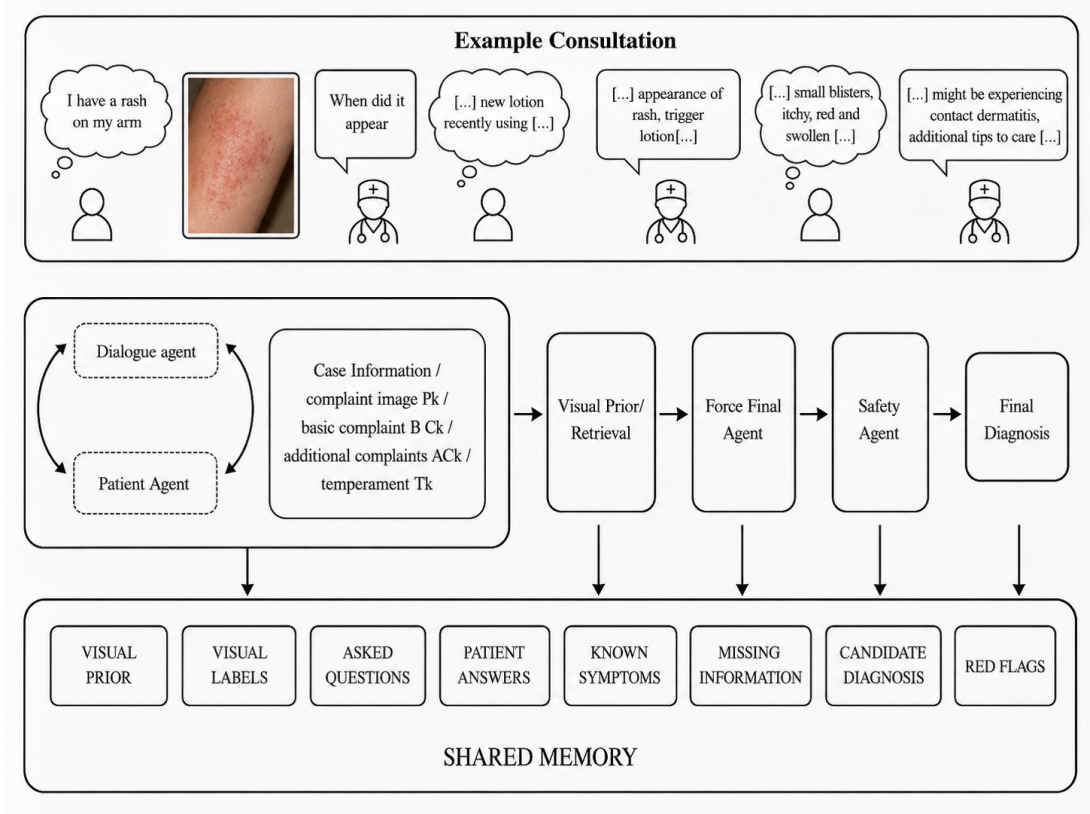


Figure 1: Overview of the proposed V-RAM framework.

Dialogue: The Doctor Agent interacts with the Patient Simulator Agent for upto $T = 4$ turns. At each turn, the agent decides whether to ask a follow-up question or commit to a diagnosis.

Commit: Every agent diagnosis is validated against $\mathcal{C}$. In case of an invalid label or turn budget exhaustion, the Force Final Diagnosis module selects label from $\mathcal{C}$, followed by the Safety Recommendation Agent wrapping the final diagnosis in a structured safe response.

3.2.1 CLIP Retrieval Index

The retrieval corpus contains cases $\{(x_i, y_i)\}_{i=1}^N$, where x_i is the medical image and $y_i \in \mathcal{L}$ is the associated ground-truth label. Each image in the corpus is encoded and L_2 normalized.

At test time, query image P_k is encoded to obtain e_q . Similarity between the query and corpus entry is given by the cosine similarity between the normalized embeddings. The set of indices of the top-k nearest neighbours are:

$$\mathcal{N}_k(I) = \{i_1, \dots, i_k \mid s_{i_1} \geq \dots \geq s_{i_k}, i_j \neq \text{idx}(I)\}. \quad (1)$$

where query case itself is excluded to avoid data leakage.

$$c(l) = |\{i \in \mathcal{N}_k(I) \mid y_i = l\}|, \quad \forall l \in \mathcal{L} \quad (2)$$

The top-3 most frequent labels among the retrieved neighbours form the retrieval cue set:

$$\mathcal{R}(I) = \{l_1, l_2, l_3 \mid c(l_1) \geq c(l_2) \geq c(l_3)\} \quad (3)$$

This set $\mathcal{R}(I)$ is stored in the structured memory and contributes to the candidate label set $\mathcal{C}$ alongside the visual prior(Section X).

3.2.2 Visual Prior Agent

The Visual prior Agent proposes an initial differential diagnosis from the image alone. For the VLM backbone, parametrized by its prompt denoted by $g: \mathcal{X} \times \mathcal{T} \rightarrow \mathcal{T}$:

$$v(I, C_0) = \{v_1, v_2, v_3 \mid v_i \in \mathcal{L}\} \quad (4)$$

The output is in a constrained format, that gives the VISUAL_TOP_N: <label>. The labels are extracted from raw text by a deterministic parser p_{parse} and a fuzzy string matching with a cutoff of 0.75 is applied to resolve inconsistencies.

$$v_i = p_{\text{parse}}(g(\text{prompt}_v \mid I, C_0))_i \quad (5)$$

Any invalid, malformed or unparseable labels are discarded. If there are less than three valid labels, the remaining candidates are omitted.

3.2.3 Structured Memory

A compact case memory M_t is maintained as a shared registry that all agents read from and update.

Memory Initialisation: At $t = 0$, memory is initialised with the basic complaint and empty fields. It is then updated with the raw visual prior v_{raw} , parsed visual candidates $\mathcal{V}$, retrieval cues $\mathcal{R}$, and retrieved-neighbour metadata ρ :

$$M_{\text{init}} = M_0 \cup \{v_{\text{raw}}, \mathcal{V}, \mathcal{R}, \rho\}. \quad (6)$$

Memory Trimming: After each turn, list fields are kept to the last $L_{\text{max}} = 6$ entries to avoid context overflow:

$$\text{trim}(M_t) = \{k \mapsto M_t[k][-L_{\text{max}}:] \mid k \in \mathcal{F}\}. \quad (7)$$

Here, $\mathcal{F}$ contains asked questions, patient answers, known symptoms, missing information, and candidate diagnoses.

3.2.4 Candidate Restriction

Before the dialogue loop begins, the Candidate Label Set is constructed by the union of Visual Prior and the retrieval cues.

$$\mathcal{C} = \{v_1, v_2, v_3\} \cup \{r_1, r_2, r_3\} \quad (8)$$

This set is written to memory and remains stable throughout the episode. Commitment to a label outside $\mathcal{C}$ is prevented. If attempted, the decision is overridden by the Force Final Diagnosis Module. This stability is important because the candidate set is grounded in visual evidence before the dialogue begins. Without this constraint, the agent may drift toward diagnoses influenced by misleading patient responses or earlier turns. The visual candidate set therefore acts as an anchor for the dialogue.

3.3 Structured Dialogue Loop

The visual candidates are fixed in memory and the dialogue loop of upto $T = 4$ turns is entered. Each turn follows the same three step pattern: decisions

are made by the Doctor Agent, responses are generated by the Patient Simulator and records are made by the Memory Update Agent.

3.3.1 Patient Simulator Agent

Realistic patient responses are generated by the Patient Simulator p , conditioned on hidden ground truth symptoms and a personality type. Let H be the set of hidden complains and $\phi \in \Phi = \{\text{sanguine, choleric, melancholic, phlegmatic}\}$ the temperament profile:

$$p(q_t, C_0, H, \phi) = a_t \quad (9)$$

The full hidden symptom list H is provided to the agent, but proper guardrails ensure it reveals only information relevant to the current question. Hidden symptoms are never volunteered. The response style is driven by the personality ϕ .

3.3.2 Doctor Agent

Decisions on the whether a follow-up question should be asked or commit to a diagnosis are made by the Doctor Agent d :

$$d(M_t, t, T) = \begin{cases} \text{ASK}(q_{t+1}), \\ \text{FINAL_DIAGNOSIS}(\hat{y}), \hat{y} \in \mathcal{L}. \end{cases} \quad (10)$$

Behavioral rules are encoded in the prompt to ensure that asked questions are not repeated, visual and retrieval labels are treated as strong evidence.

Memory Update Agent: After each (q_t, a_t) pair, the Memory Update Agent μ converts the dialogue into structured updates: new symptoms $\Delta\Sigma_t$, missing information $\Delta\Delta_t$, and red-flag indicators $c_t, p_t \in \{\text{yes, no}\}$:

$$\mu(M_{t-1}, q_t, a_t) = (\Delta\Sigma_t, \Delta\Delta_t, c_t, p_t). \quad (11)$$

Here, $\Delta\Sigma_t$ denotes newly reported symptoms, $\Delta\Delta_t$ denotes unresolved clinical information, and c_t, p_t indicate whether red flags were checked or present. The memory is updated as:

$$M_t[\text{known-symptoms}] = M_{t-1}[\text{known-symptoms}] \cup \Delta\Sigma_t. \quad (12)$$

$$M_t[\text{missing-info}] = \Delta\Delta_t. \quad (13)$$

3.4 Safety and Fallback

After a diagnosis $\hat{y}$ is produced by the dialogue loop, validity and safety of the response has to be ensured.

3.4.1 Force Final Diagnosis

The Force Final Diagnosis Module F can be invoked when any one of the three conditions are true. 1. $\hat{y} \notin \mathcal{C}$ or agent predicts outside label-space, 2. $t = T$ without a commitment or turn budget exhaustion 3. $\hat{y} = \text{unknown}$

The candidate set $\mathcal{C}$, the initial complaint C_0 , and the accumulated memory M_t are presented to the VLM by a dedicated prompt, with strict instructions to choose only from $\mathcal{C}$:

$$F(I, C_0, M_t, \mathcal{C}) = \hat{y} \in \mathcal{C} \quad (14)$$

If an invalid label is still returned by the VLM, a deterministic fallback applies majority vote over $\mathcal{C}$:

$$\hat{y}_{\text{fallback}} = \arg \max_{l \in \mathcal{C}} \sum_{l' \in \mathcal{C}} [l = l'] \quad (15)$$

Ties are broken by picking the first label in $\mathcal{C}$ sorted alphabetically.

3.4.2 Safety Recommendation Agent

The fixed diagnosis $\hat{y}$ is molded into a safe response by the Safety Recommendation Agent s without modifying the label.

$$s(\hat{y}, C_0, v_{\text{raw}}, M_t) = (\text{REASON}, \text{UNCERTAINTY}, \text{RECOMMENDATION}) \quad (16)$$

Uncertainty $u \in \{\text{low}, \text{medium}, \text{high}\}$ and the recommendation includes guidance on when to seek in-person care. The agent is also forbidden from suggesting prescription medicine. Quantitative metrics are computed solely on $\hat{y}$.

3.5 Dataset

The 3MD-Bench dataset (Sviridov et al., 2025) is used under the MIT License, which contains 2996 test cases across 34 disease categories, spanning five medical domains. Each sample includes a medical image P_k , a basic complaint C_0 , hidden additional complaints H , patient temperament ϕ and ground truth diagnosis y .

3.6 Evaluation Metrics

Exact Accuracy, Macro F1, Micro F1, and Weighted F1 are reported (Pedregosa et al., 2011). Macro F1 is used as the primary metric to account for class imbalance across the 34 disease categories. All metrics are computed with exact string matching after lower-casing and whitespace normalization.

3.7 Experimental Setup

The full multi-agent system is evaluated across four backbone VLMs: GPT-4o-mini (OpenAI, 2024) via the OpenAI API, Llama 3.2 Vision 11B Instruct, MedGemma 4B it (Sellergren et al., 2025) and Qwen 2 VL 2B Instruct (Wang et al., 2024) via HuggingFace. The entire evaluation across the open source models was done on an RTX Pro 6000 Server Edition via Lightning AI, running for 40 GPU-hours.

The hyper-parameters MAX_TURNS, temperature and token budget chosen for the model were determined empirically using the Optuna library (Akiba et al., 2019). were chosen as hyper-parameters and tuned.

4 Results

As mentioned earlier, the proposed framework is compared with the baseline, containing dialogue and images, and for GPT-4o-mini, an additional baseline with rationale. The macro F1 and the number of utterances obtained from the methods are presented in Table 1 and the accuracy, precision and recall scores attained by the different methods are provided in Figure 2. It should be noted that the paper includes representative samples of the evaluation metrics computed from the results obtained using the proposed methods, along with an ablation study, to ensure better continuity and readability for the reader. Additional metrics are provided in the appendix (refer 6).

It is observed from the table that GPT-4o-mini achieves the highest scores across all the metrics. Among the open-weight models, performance is positively correlated to the model scale. We also see that our approach gives us significant gain as compared to other LVLMs like Qwen, Llama 3.2 and MedGemma. The addition of rationale-aware reasoning yields a gain of 2.2 macro F1 points for GPT-4o-mini, whereas we notice a slight degradation for Qwen 2 VL 2B Instruct. This indicates that smaller models are less able to accommodate

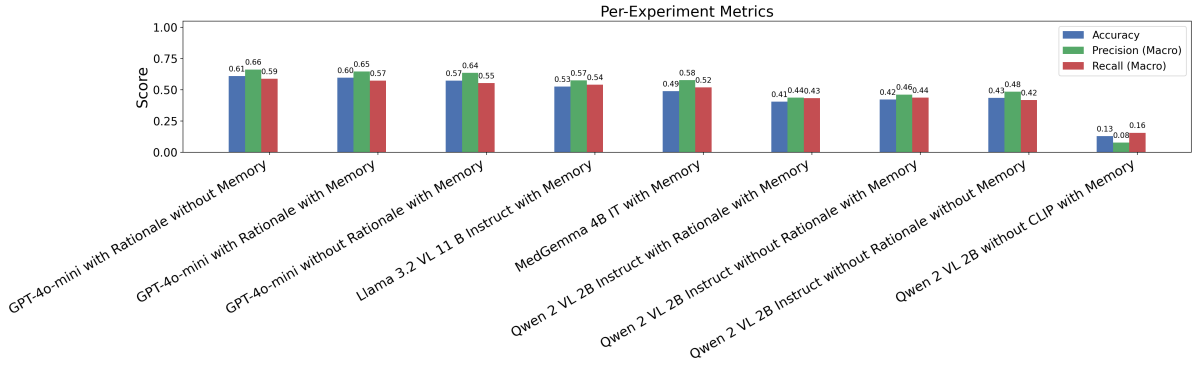


Figure 2: Macro Precision, Macro Recall and Accuracy across the different experimental configurations for different VLM backbones on 3MDBench

Table 1: Comparison of V-RAM with the Baseline across different VLM backbones

Model	Configuration	Macro F1	Number of Utterances
GPT-4o-mini	CLIP + rationale, no memory (ours)	59.54	2.84 ± 0.79
	CLIP + rationale + memory (ours)	58.47	2.74 ± 0.74
	CLIP + memory, no rationale (ours)	56.25	2.63 ± 0.65
	dialogue + image + rationale (baseline)	56.9	14.99 ± 4.23
Llama 3.2 Vision 11B	CLIP + memory (ours)	52.47	3.93 ± 0.35
	dialogue + image (baseline)	41.5	14.49 ± 4.02
MedGemma 4B	CLIP + memory (ours)	51.07	3.98 ± 0.18
	dialogue + image (baseline)	37.9	17.48 ± 4.84
Qwen2-VL-2B	CLIP + memory (ours)	41.53	3.98 ± 0.19
	CLIP + memory + rationale(ours)	40.60	3.4 ± 1.2
	dialogue + image (baseline)	39.0	15.11 ± 4.39

the additional reasoning overhead within the context window and fixed token budget, while GPT-4o-mini benefits from the 4x larger 128k token context window in comparison to Qwen 2 VL 2B’s 32k.

Additionally, the gain in performance is reported for maximum number of utterances as 4, roughly a 76% reduction from the maximum 28, as proposed by Sviridov et al. (Sviridov et al., 2025).

To interpret the visual manifestations of the symptom driving the CLIP-based retrieval, attention heatmaps were generated for each query using the ViT-B/32 encoder. Attention rollout (Abnar and Zuidema, 2020) was applied and gradient weighted class activation maps were computed by back-propagating the image-text similarity score (Selvaraju et al., 2017) shifts to off-topic regions for wrong label and this corroborates that the retrieval module operates on pathology-grounded visual features rather than spurious correlations.

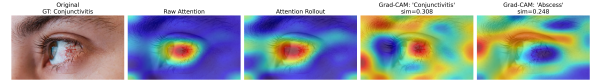


Figure 3: Grad-CAM heatmaps for CLIP ViT-B/32 conditioned on different diagnosis prompts. When prompted with the ground-truth label (left), the heatmap concentrates on the clinically relevant region. Prompting with an incorrect label shifts the activation to diffuse or off-target areas, demonstrating that the cross-modal similarity score captures pathology-grounded visual features rather than spurious correlates.

4.1 Ablations

Three ablations are performed, all at $T = 4$ turns. Firstly, CLIP retrieval is disabled to allow the candidate set to be seeded by the Visual Prior Agent alone. Secondly, structured memory is disabled to allow the dialogue agent to operate directly on the raw transcript. Finally, model is prevented from verbalizing CoT (Wei et al., 2022), allowing only implicit understanding to drive the diagnosis. The ablations are performed for GPT-4o-mini and Qwen 2 VL 2B Instruct specifically to understand

Table 2: Ablation study on Qwen2-VL-2B-Instruct

Configuration	Acc.	Macro F1	Wt. F1
Qwen + CLIP + Memory	42.19	41.53	40.57
Qwen + Memory	12.82	8.27	7.75
Qwen + CLIP	43.49	40.34	39.23

the efficacy of the different modules. The result obtained for the ablations are presented in Table 2 and disease wise metrics are provided in the 6.

Role of Retrieval and Memory: *Without CLIP*, the performance drops to near random levels, bearing a macro F1 of 8.27, with random expectation at approximately at 3% for 34 classes. The Visual Prior Agent alone, lacks retrieval-augmented grounding and is insufficient to reliably propose the correct diagnosis hypothesis using only the VLM’s parametric knowledge gained during training. This confirms that zero-shot CLIP index provides critical visual prior that cannot be extracted from the image by the VLM alone.

Without Structured Memory, accuracy is slightly higher while the macro F1 is reduced. The slight accuracy improvement is attributed to more effective context being available for the majority class predictions when the memory updates is removed, which frees prompt space by reducing structured field overhead. The subsequent macro F1 decrease indicates that the class balancing signal from memory is necessary for maintaining consistent performance across all 34 categories. Without memory, majority classes are over-predicted at the cost of minority categories.

Role of Structured Memory with verbalised Chain-of-Thought:

On the stronger GPT-4o-mini with verbalised CoT, both accuracy and macro F1 are improved when memory is removed. This confirms that the role of structured memory is partially subsumed by the verbalised reasoning in the rationale-aware agents when known symptoms, missing information, and diagnostic hypothesis are tracked by the model within its own CoT. The explicit memory agent, adds redundant structure that dilutes the prompt space signal without commensurate benefit. Across these ablations, three findings were seen: (1) CLIP retrieval is essential for visual grounding, to avoid near random performance. (2) structured memory provides complementary class balancing signal that is more valu-

able for smaller models and minority classes. (3) On larger models, with verbalised CoT, the memory function is internalised, rendering the explicit formatted memory agent redundant.

5 Limitations

The system relies entirely on visual similarity patterns captured by CLIP and does not incorporate external medical knowledge bases, clinical guidelines, or predefined set of rules. While this makes the system training-free and generalizable, it may miss rare conditions or atypical presentations that require medical knowledge.

There are some conditions often present with visually similar manifestations, making differential diagnosis challenging. The system relies entirely on visual similarity patterns captured by CLIP. However, in some conditions such as eczema, contact dermatitis often exhibit overlapping visual features. Consequently, a definitive diagnosis frequently requires additional clinical context, patient history, patch testing, or biopsy information that is not available in the current benchmark setting. Hence, the system’s diagnostic accuracy is inherently constrained by this clinical ambiguity, and real-world deployment would require oversight by qualified dermatologists or healthcare professionals to handle complex or ambiguous cases.

6 Conclusion

V-RAM demonstrates that advancements in the architectural framework by adding a visual retrieval, structured memory and multi agent reasoning can substantially improve diagnostic performance of vision-language models without requiring expensive fine-tuning. The proposed method remains training-free and can be used with different LVLM backbones

The results obtained from the proposed method elicit the rationale of using visual prior along with dialogue can enhance the efficiency of disease diagnosis systems. The ablation study results reveal that structured memory contributes most significantly to performance, followed by visual retrieval, suggesting that explicit context management matters more than raw model capacity for multi-turn diagnostic reasoning.

These findings challenge the prevailing assumption that medical AI systems must rely on either massive proprietary models or extensive domain-

specific fine-tuning. Instead, V-RAM shows that a system design combining retrieval and agent specialization can unlock strong performance. However, it has been observed that ambiguity in the visual priors can degrade the performance of the model. The future work will focus on refining visual prior disambiguation mechanisms and extending the multi-agent framework to complex, multi-specialty clinical workflows.

Use of AI Assistance

AI-based writing tools were used only for grammar correction, language polishing, and improving readability. The scientific content, methodology, experiments, analysis, and conclusions were developed and verified by the authors. No AI tool was used to generate experimental results, modify data, or make scientific claims without author review.

Code Availability

All code, prompts, evaluation scripts, and configuration files required to reproduce the experiments are provided in an anonymized GitHub repository. The repository link is included in the supplementary material to preserve double-blind review anonymity.

References

- Ivan Sviridov, Amina Miftakhova, Artemiy Tereshchenko, Galina Zubkova, Pavel Blinov, and Andrey Savchenko. 3MDBench: Medical multimodal multi-agent dialogue benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 26614–26654, 2025.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.
- Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4190–4197, 2020.
- Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.
- Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631, 2019.
- OpenAI. GPT-4o system card. *OpenAI Technical Report*, 2024. <https://openai.com/index/gpt-4o-system-card/>
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, 2019.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning (CHIL)*, pages 248–260, 2022.
- Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. MediQ: Question-asking LLMs and a benchmark for reliable interactive clinical reasoning. *arXiv preprint arXiv:2406.00922*, 2024.
- Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hongru Wang, Yuhang Guo, Tong Ruan, Jie Xu, Xiaofan Zhang, and Shaoting Zhang. Medical dialogue system: A survey of categories, methods, evaluation and challenges. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2840–2861, 2024.
- Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I. Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. Guidelines for rigorous evaluation of clinical LLMs for conversational reasoning. *arXiv preprint arXiv:2401.04547*, 2024.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474, 2020.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Papatat, and Ming-Wei Chang. REALM: Retrieval-augmented language model pre-training. In *Pro-*

- ceedings of the 37th International Conference on Machine Learning (ICML)*, pages 3929–3938, 2020.
- Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 874–880, 2021.
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.
- Wenqi Xiong, Bingsheng Chen, Shujian Wang, and Shicheng Tan. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*, 2024.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.
- Janet L. Kolodner. An introduction to case-based reasoning. *Artificial Intelligence Review*, 6(1):3–34, 1992.
- Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. MDAgents: An adaptive collaboration of LLMs for medical decision-making. *arXiv preprint arXiv:2404.15155*, 2024.
- Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agent-Clinic: A multimodal agent benchmark to evaluate AI in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.
- Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *arXiv preprint arXiv:2402.09742*, 2024.
- Jianing Qiu, Kyle Lam, Guohao Li, Amish Acharya, Tien Yin Wong, Ara Darzi, Wu Yuan, and Eric J. Topol. LLM-based agentic systems in medicine and healthcare. *Nature Machine Intelligence*, 6(12):1418–1420, 2024.
- Mohammad Almansoori, Komal Kumar, and Hisham Cholakkal. Self-evolving multi-agent simulations for realistic clinical interactions. *arXiv preprint arXiv:2503.22678*, 2025.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837, 2022.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.
- Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 22199–22213, 2022.
- Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pages 19724–19731, 2024.
- Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023.
- Chenxu Hu, Jie Fu, Chenzhuang Du, Guande He, and Hang Zhao. ChatDB: Augmenting LLMs with databases as their symbolic memory. *arXiv preprint arXiv:2306.03901*, 2023.
- Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. MedAlpaca—an open-source collection of medical conversational AI models and training data. *arXiv preprint arXiv:2304.08247*, 2023.
- Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Subramanian, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023.
- Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. Towards conversational diagnostic artificial intelligence. *Nature*, 642:442–450, 2025.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keyang Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cián Hughes, Charles Lau, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1820–1828, 2021.

Abbi Ward, Jimmy Li, Julie Wang, Sriram Lakshminarasimhan, Ashley Carrick, Bilson Campana, Jay Hartford, Pradeep K. Sreenivasaiah, Tiya Tiyasirisokchai, Sunny Virmani, et al. Creating an empirical dermatology dataset through crowdsourcing with web search advertisements. *JAMA Network Open*, 7(11):e2446615, 2024.

A Disease Wise Model Performance

Code Availability

All code, prompts, evaluation scripts, and configuration files required to reproduce the experiments are available in an anonymized repository: <https://anonymous.4open.science/r/V-RAM-7C0E/>.

This section provides a more detailed disease-level analysis of the proposed V-RAM framework across all evaluated system configurations. While the main paper reports aggregate metrics such as accuracy and macro F1, disease-wise scores help reveal which clinical categories benefit most from visual retrieval, structured memory, and multi-agent reasoning. The confusion matrix shows that performance varies across disease classes, reflecting the difficulty of distinguishing visually similar conditions and the effect of class imbalance in the benchmark. These results provide a finer-grained view of model behavior beyond the average scores reported in the main results section.



Figure 4: Per-disease Macro F1 scores (%) across all nine system configurations.

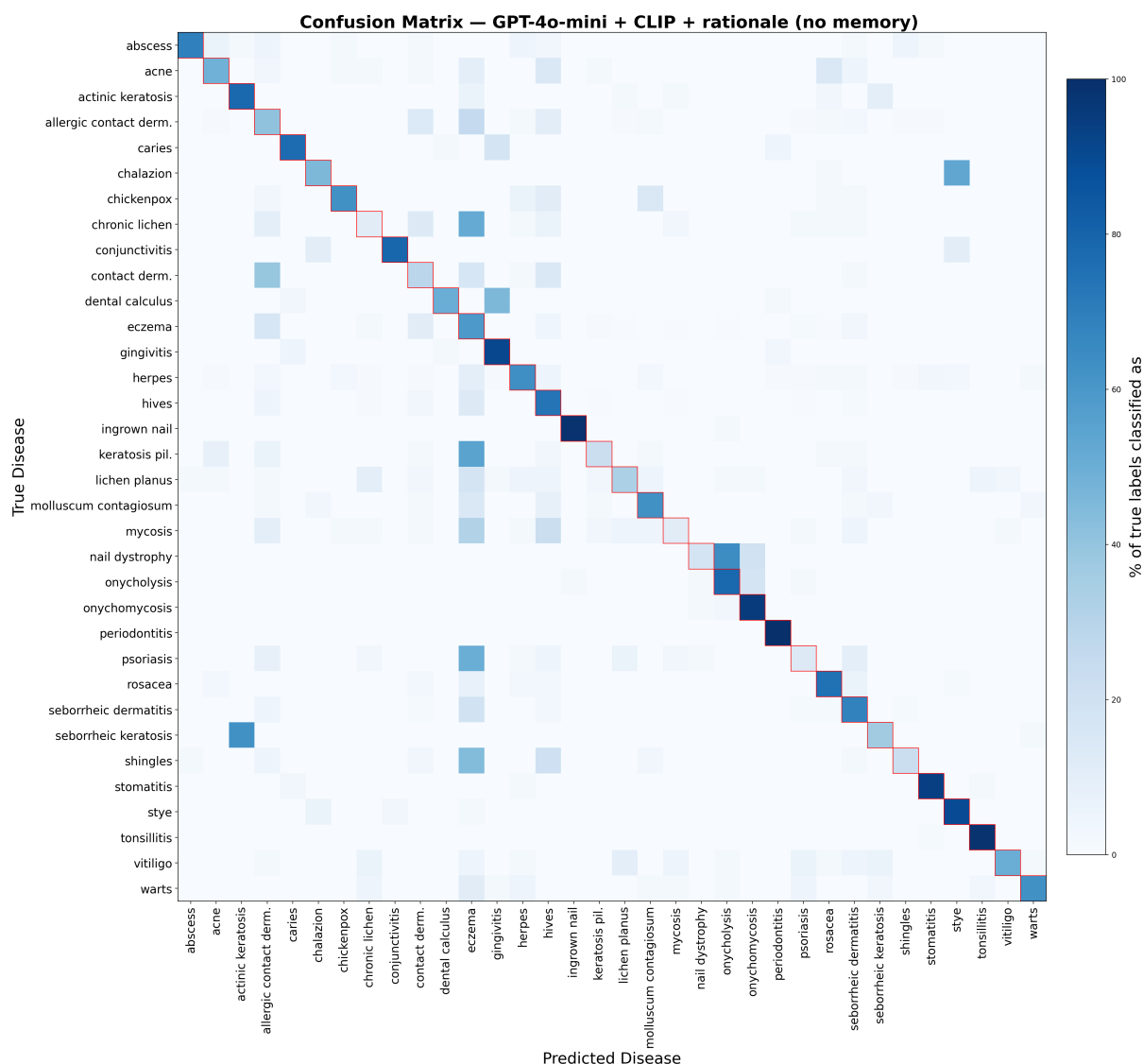


Figure 5: Confusion matrix for the best-performing system (GPT-4o-mini + CLIP + rationale, no memory; Macro F1: 59.54). Rows correspond to true disease labels and columns to predicted labels; values are row-normalized to show the percentage of true cases classified as each predicted disease. The red diagonal highlights correct predictions. Off-diagonal entries reveal the most common confusions: allergic contact dermatitis and eczema are frequently mistaken for one another, and keratosis pilaris is often confused with acne and eczema.

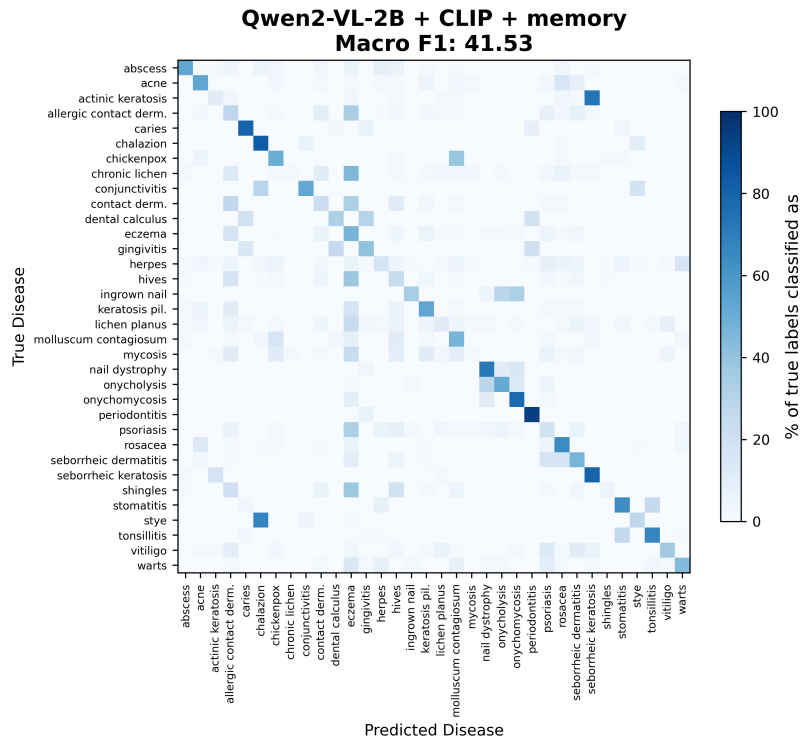


Figure 6: Confusion matrix for Qwen2-VL-2B with CLIP + memory (41.53 F1).

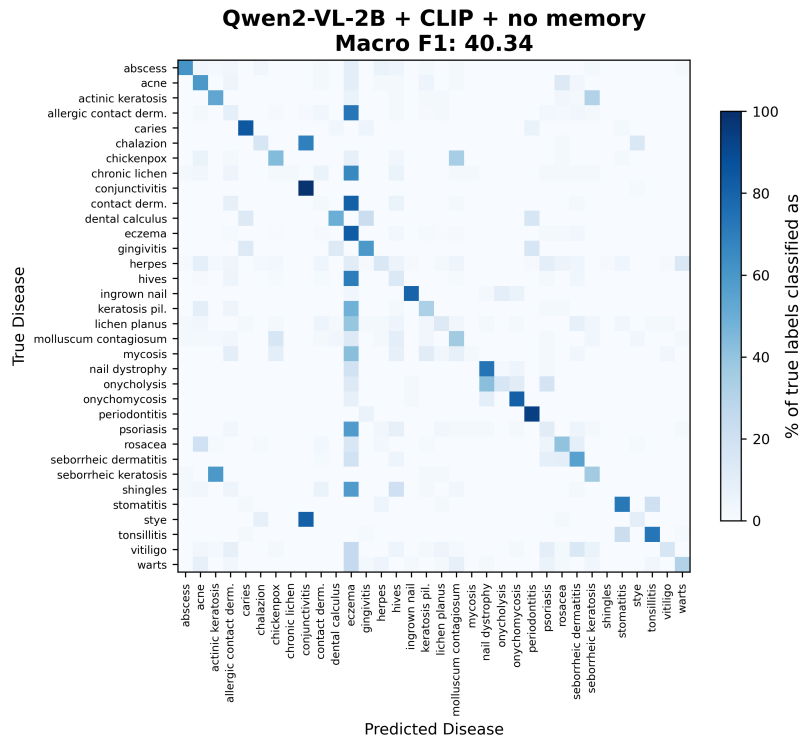


Figure 7: Confusion matrix for Qwen2-VL-2B with CLIP, no memory (40.34 F1).

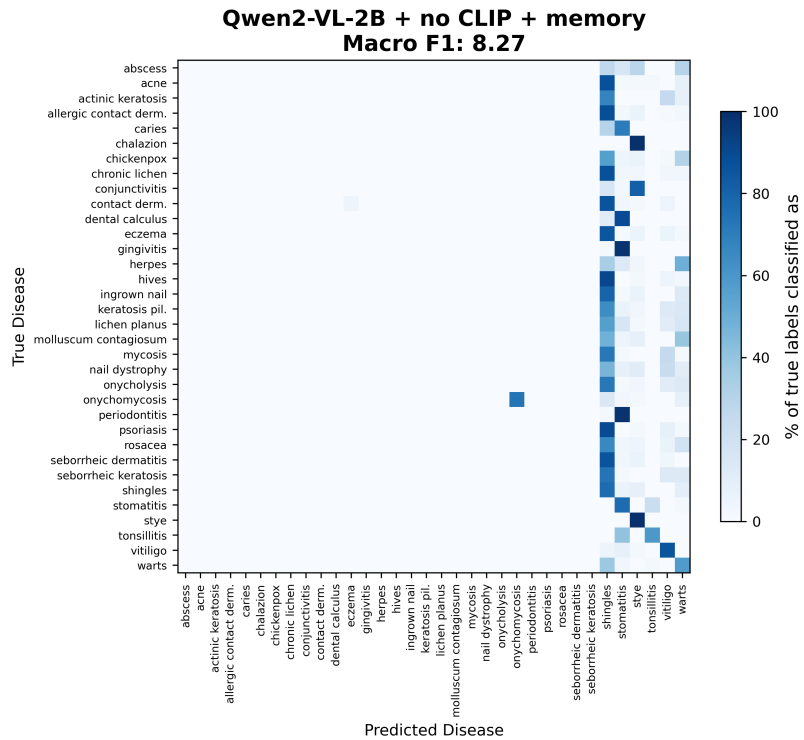


Figure 8: Confusion matrix for Qwen2-VL-2B with no CLIP + memory (8.27 F1). Removing CLIP retrieval collapses predictions into a small set of frequent labels.

B Stability of Responses Across Models

This section analyzes the stability of diagnostic decisions across different model configurations. We measure cross-model disagreement for each disease class using entropy, where higher entropy indicates that different models produce less consistent predictions. This helps identify diseases for which the diagnosis is more sensitive to the choice of backbone model or framework configuration. Stable classes suggest stronger visual or textual signals, whereas high-entropy classes may require additional clinical context, better retrieval cues, or stronger reasoning mechanisms.

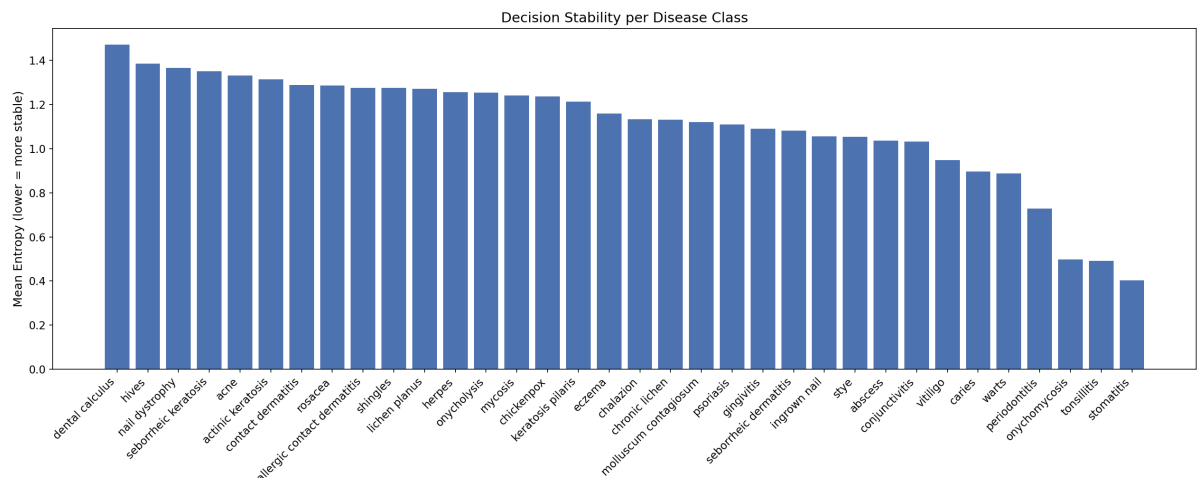


Figure 9: Decision stability per disease class measured by mean cross-model entropy across all nine system configurations. Higher entropy indicates greater disagreement among models for that disease.

C Prompts

C.1 Patient Simulator Agent

Patient simulator prompt

You are a simulated patient in a medical consultation.

Basic complaint:

[general_complaint](#)

Hidden/additional symptoms:

[hidden_text](#)

Personality style:

[personality](#)

Rules:

- Answer only the doctor's current question.
- Do not reveal all hidden symptoms at once.
- Do not mention the diagnosis.
- Keep the answer short and natural.
- If the doctor asks something related to a hidden symptom, reveal only that relevant detail.
- If the doctor asks something unrelated, answer briefly.

Doctor question:

[doctor_question](#)

Patient answer:

C.2 Visual Prior Agent

Visual Prior Agent prompt

You are the Visual Prior Agent. Look at the medical image and the basic complaint. Choose the top 3 most likely diagnoses from the allowed labels.

Allowed labels:

[DISEASE_LIST](#)

Basic complaint:

[general_complaint](#)

Rules:

- Use only labels from the allowed list.
- Do not explain.
- Do not add extra text.
- Output exactly this format:

VISUAL_TOP_1: <label>

VISUAL_TOP_2: <label>

VISUAL_TOP_3: <label>

C.3 Memory Update Agent

Memory Update Agent prompt

You are the Memory Update Agent. Update the structured medical memory based on the latest doctor question and patient answer.

Current memory:

`memory_to_text(memory)`

Latest doctor question:

`latest_question`

Latest patient answer:

`latest_patient_answer`

Return only this format:

KNOWN_SYMPTOMS: <comma-separated short symptoms or none>

MISSING_INFO: <comma-separated important missing clinical details or none>

RED_FLAGS_CHECKED: <yes/no>

RED_FLAGS_PRESENT: <yes/no>

C.4 Doctor Agent

Doctor Agent prompt

You are the Doctor Agent. Your job is to decide whether to ask one more question or give the final diagnosis. First, think through the case internally. Then give your answer.

Allowed labels:

`DISEASE_LIST`

Structured memory:

`memory_to_text(memory)`

Your internal reasoning from previous turns:

`doctor_internal_history`

Current turn: `turn` **Maximum turns:** `max_turns`

Rules:

- First, reason internally about what you know, what is missing, and what to do next.
- Then output your action.
- Do not repeat any question already asked.
- Use known symptoms from memory.
- Use missing information from memory to decide the next question.
- Treat visual top labels and retrieved visual labels as strong evidence.
- If enough information is available, output exactly one diagnosis label.
- On the final turn, you **MUST** give a final diagnosis.
- Use only labels from the allowed list.

Output exactly this format:

REASONING: <your internal diagnostic reasoning>

ANSWER: ASK: <one follow-up question>

or

ANSWER: FINAL_DIAGNOSIS: <label>

Example (asking):

REASONING: Red scaly patches on elbows; patient reports night itching; missing joint pain info.

ANSWER: ASK: Have you experienced any joint pain or swelling?

Example (diagnosing):

REASONING: Well-demarcated red plaques with silvery scale; family history of psoriasis; no red flags.

ANSWER: FINAL_DIAGNOSIS: psoriasis

C.5 Final Diagnosis Agent

Final Diagnosis Agent prompt

Choose exactly one final diagnosis from the candidate labels only.

Candidate labels:

[candidate_list](#)

Basic complaint:

[general_complaint](#)

Structured memory:

[memory_to_text\(memory\)](#)

Rules:

- Choose only from the candidate labels.
- Prefer the most specific label supported by image, retrieved visual labels, and patient symptoms.
- Do not choose a generic label if a more specific candidate fits.
- Do not explain.
- Output exactly:

FINAL_DIAGNOSIS: <label>

C.6 Safety Recommendation Agent

Safety Recommendation Agent prompt

You are the Safety Recommendation Agent. The diagnosis has already been selected. You are **NOT** allowed to change the diagnosis.

Fixed diagnosis:

[fixed_diagnosis](#)

Basic complaint:

[general_complaint](#)

Visual prior:

[visual_prior](#)

Dialogue history:

[dialogue_history](#)

Write a short safe response in this format:

FINAL_DIAGNOSIS: [fixed_diagnosis](#)

REASON: <brief reason based on image and symptoms>

UNCERTAINTY: <low/medium/high>

RECOMMENDATION: <safe advice and when to see a clinician>

Rules:

- Do not change FINAL_DIAGNOSIS.
- Do not suggest prescription medication.
- Mention seeing a clinician if symptoms worsen or red flags appear.

D Sample Case Examples

This section provides representative examples of V-RAM operating on individual medical dialogue cases. Each example includes the patient–doctor interaction.

D.1 Sample Case Example on V-RAM

Sample Case

Patient: I have dry and itchy skin.

Doctor: Can you tell me how long you have been experiencing these symptoms?

Patient: About two weeks.

Doctor: Have you started using any new skin products, soaps, or detergents during this time?

Patient: No, I haven't made any changes to my skincare routine or used any new products.

Doctor: Have you noticed any additional symptoms, such as redness, swelling, or weeping from the affected areas?

Patient: Yes, the affected areas are slightly red and inflamed. And when I scratch, I sometimes notice oozing.

Doctor: Based on the symptoms you've described, it seems you may be experiencing eczema. I recommend keeping the affected areas moisturized, avoiding scratching, and considering an over-the-counter hydrocortisone cream to help with inflammation. If the condition worsens or does not improve, please consult a healthcare professional for further evaluation and treatment.

Patient: I want to know more about the hydrocortisone cream. How quickly can I expect to see results?

Doctor: Typically, you may start to see improvement within a few days of applying hydrocortisone cream, especially for mild to moderate cases. However, for best results, it's important to use it as directed—usually applying it twice daily. If you don't see improvement within a week, or if your symptoms worsen, I recommend consulting a healthcare professional for further evaluation.

Patient: That sounds reasonable. What about the sensitivity to certain soaps and detergents? Should I avoid using those altogether?

Doctor: It may be beneficial to avoid using soaps and detergents with strong fragrances or harsh chemicals. Instead, consider using gentle, fragrance-free products specifically formulated for sensitive skin. It's also a good idea to perform a patch test with any new product before full use. If you notice irritation after using a product, it's best to discontinue it.

Patient: I'd like to get rid of this itchiness as soon as possible. Can I use the hydrocortisone cream on the raised bumps as well?

Doctor: Yes, you can use hydrocortisone cream on the raised bumps to help relieve itchiness and reduce inflammation. Just ensure to apply it gently and avoid overusing it, as prolonged use can lead to skin thinning. Follow the instructions on the product for best results. If the itching persists or worsens, consult a healthcare professional.

Patient: I'll make sure to use it as directed. Thanks for the advice.

Doctor: You're welcome! If you have any more questions or if your symptoms don't improve, feel free to reach out. Take care!



Case ID: 1233

Ground Truth: allergic contact dermatitis

Predicted: allergic contact dermatitis

Visual Top-3:

- eczema
- contact dermatitis
- allergic contact dermatitis

Final Output:

FINAL_DIAGNOSIS: allergic contact dermatitis

REASON: The symptoms of redness, swelling, itching, and the presence of blisters suggest an allergic reaction to a new product, consistent with allergic contact dermatitis.

UNCERTAINTY: low

RECOMMENDATION: Avoid using the new lotion and any other potential irritants. Keep the affected area clean and moisturized. If symptoms worsen, or if you experience severe swelling, difficulty breathing, or signs of infection, seek medical care promptly.

D.2 Sample Case Example V-RAM: Conjunctivitis

Sample Case

Patient: I feel discomfort in my eye.

Doctor: Are you experiencing any redness, swelling, or discharge from your eye?

Patient: Yes, my eye is quite red, and there is some swelling around it.

Doctor: Is there any discharge from your eye?

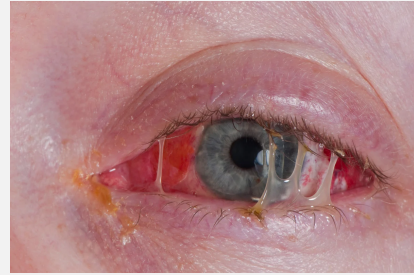
Patient: Yes, I've noticed some increase in tearing and a bit of discharge.

Doctor: Is the discharge watery or thick and yellow/green?

Patient: It's mostly watery, but there is some yellowish discharge as well.

Doctor: Based on your symptoms of redness, swelling, and watery discharge with a slight yellowish tint, you are most likely experiencing conjunctivitis. Keep the eye clean, avoid touching it, and consult a doctor if symptoms worsen.

Patient: Thank you, that makes sense.



Case ID: 26

Ground Truth: conjunctivitis

Predicted: conjunctivitis

Visual Top-3:

- conjunctivitis
- sty
- chalazion

Final Output:

FINAL_DIAGNOSIS : conjunctivitis

REASON : Redness, swelling, and watery discharge with yellowish tint are consistent with conjunctivitis, supported by visual and retrieval evidence.

UNCERTAINTY : low

RECOMMENDATION : Keep the eye clean and avoid rubbing. Use lubricating eye drops. If discharge becomes thick or vision is affected, seek medical care promptly.