Lightweight Joint Optimization of General-Purpose Vision-Language Models and Retrievers for RAG-Based Medical Diagnosis

Anonymous ACL submission

Abstract

Retrieving relevant visual and textual information from medical literature and hospital records can enhance diagnostic accuracy for clinical image interpretation. We develop a multimodal retrieval model jointly optimized with an LVLM for medical diagnosis, unlike standard RAG which doesn't backpropagate LVLM errors to the retriever. Using only general-purpose backbones with lightweight fine-tuning, our model achieves competitive results with medically-pretrained models on clinical classification and VQA tasks. In a novel analysis, we find that different top-retrieved images often yield different predictions for the same target, and that these cases are challenging for all models, even for non-retrieval models. Our joint retrieval optimization significantly improves these cases over standard RAG. However, oracle analysis reveals that while the correct diagnosis is frequently achievable using one of the top retrieved images, in practice there is a large performance gap from the oracle, and rerankers using frontier LVLMs do not close this gap—leaving ample room for improvement by future methods.[1]

1 Introduction

Inferring diagnoses from medical imagery is a fundamental part of clinical decision-making. Large Vision Language Models (LVLMs) have been widely explored for medical diagnosis (Thawakar et al., 2024; Wu et al., 2023; Zhang et al., 2023b; Moor et al., 2023; Li et al., 2023). To improve the performance of LVLMs in the medical field, retrieval augmentation (RAG) has been adopted and has shown promising results, providing more accurate and also explainable methods (He et al., 2024a; Xia et al., 2024b,a; Wu et al., 2025).

In this work, we explore a lightweight fine-tuning approach for taking a *general*-purpose LVLM and optimizing it together with a *general*-purpose multimodal retriever for medical diagnosis tasks. Different from the common multimodal RAG setup, we perform joint optimization of the LVLM and retriever. This teaches the LVLM to utilize retrieved knowledge—images, captions and clinical reports from both literature and hospital records—while at the same time tuning the retriever to find information that leads to correct predictions by the LVLM. Poorly optimized retrieval mechanisms can mislead models (Yoran et al., 2023; Sun et al., 2024); as shown in Figure 2, a RAG model incorrectly diagnoses a benign ultrasound when using a cancerous retrieved image, but joint optimization leads to correct retrieval and prediction.

Our method involves sequential multimodal training with a dual-head retriever architecture and

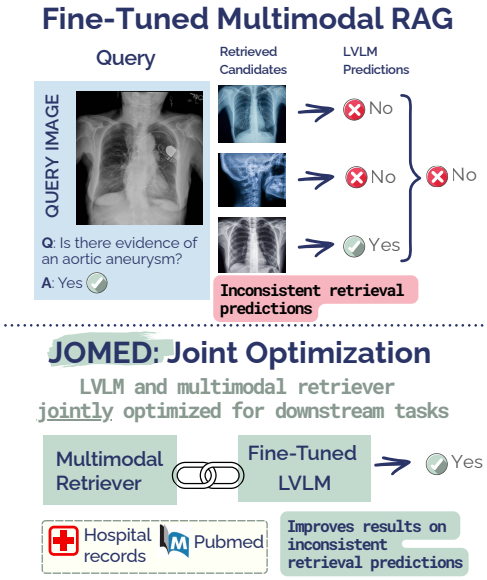

Figure 1: We jointly optimize a multimodal retriever and a Large Vision-Language Model for medical tasks. We achieve competitive results without resource-intensive medical pre-training and significantly improve performance on challenging cases where different retrieved images lead to inconsistent retrieval prediction.

[1]Code and models available at [redacted for anonymity].

a customized retrieval loss and visual question answering (VQA) training recipe. Our method achieves competitive results in medical image classification and VQA. Importantly, our focus in this work is not to chase SOTA results but to shed light on several interesting observations. One, that a simple and effective joint LVLM-retriever optimization mechanism can lead to a substantial boost in results over current multimodal RAG methods and should thus be more widely adopted for medical diagnosis. Second, that this mechanism achieves competitive results by using models with no medical pre-training (for neither the LVLM nor the retriever) and only lightweight fine-tuning. This resonates with recent findings showing general-purpose LVLMs can rival their medical counterparts (Jeong et al., 2024) and has potentially important implications considering the resource-intensive nature of pre-training processes.

Third, previous work on joint optimization in the general domain required large-scale pre-training followed by few-shot task-specific adaptation (Izacard et al., 2023; Hu et al., 2023; Lin et al., 2024). In contrast, we conduct joint optimization *directly on downstream tasks* with no pretraining and only lightweight fine-tuning (as few as 546 samples). This provides first evidence about the feasibility and utility of lightweight data-efficient joint optimization directly on downstream tasks as opposed to in the pre-training setting. We are also the first to explore and show the utility of joint optimization in the multimodal medical domain.

Fourth and importantly, we conduct a novel analysis that finds that our method helps in particular to address a class of errors we term *inconsistent retrieval predictions*. Inconsistent retrieval predictions are cases in which for a given patient image, each retrieved image leads the model to make different predictions. These cases commonly occur across our experiments and are substantially more difficult for models, both retrieval-augmented and non-retrieval models. This instability with respect to different retrieval candidates leads to high prediction entropy/uncertainty across classes, degrading not only overall RAG results but also their reliability (Lambert et al., 2022). As we show, our joint optimization mechanism significantly mitigates this issue and achieves large improvement over standard fine-tuned RAG on these cases. We further discover that for a large proportion of these inconsistent cases, at least one retrieved candidate enables the model to predict the correct answer in an oracle setting in which we provide the model with the correct retrieved candidate. This suggests useful information is being retrieved but is often lost among less helpful candidates. To address this, we explored using the powerful multimodal model o3 (OpenAI, 2025) as a reranker to identify the most predictive instances. We found it did not close the performance gap to the oracle and was comparable or inferior to our method, leaving significant room for future improvement.

Our contributions are three-fold:

- We jointly optimize a multimodal retriever and an LVLM for medical classification and VQA. We use only general-purpose backbones without any medical pretraining, and show that by retrieving medical information and co-optimizing the retriever and LVLM we can close much of the gap to expensive medically pre-trained models.

- Unlike previous general-domain joint optimization that required pre-training, we conduct joint optimization directly on downstream tasks with only lightweight data-efficient fine-tuning.

- We identify *inconsistent retrieval predictions*, challenging cases where different retrieved images lead to different predictions for the same query. Our joint optimization method significantly improves performance on these cases when compared to standard RAG.

2 Related Work

Retrieval Augmented joint optimization has been explored primarily in unimodal (text only) NLP with work such as ATLAS (Izacard et al., 2023), which conducted large-scale pretraining of both reader and retriever models and evaluated in zero-shot and few-shot scenarios in the general domain. REVEAL (Hu et al., 2023) extended this to the multimodal setting, achieving state-of-the-art results in visual tasks in the general domain with extensive pre-training. Shi et al. (2023) proposed training textual retrievers based on reader performance, influencing subsequent joint-optimization methods. Lin et al. (2024) introduced another joint optimization variant and proposed sequential training: first training the reader with retrieval augmentation, then training the retriever. Siriwardhana et al. (2023) demonstrated joint optimization's effectiveness for

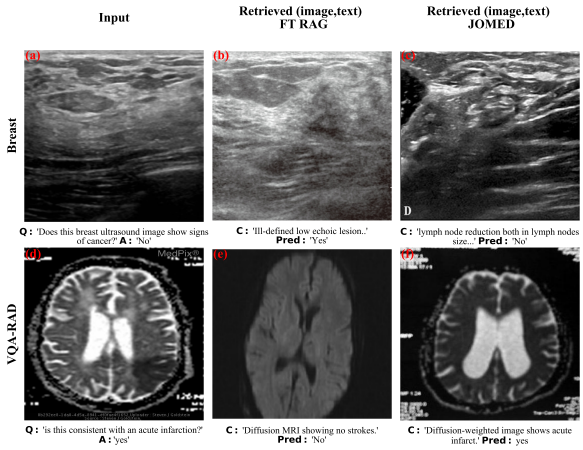

Figure 2: joint optimization impacts inconsistent retrieval predictions. After joint optimization, retrieved images show greater alignment with query image labels. In Breast, a cancer-free ultrasound query (a) initially retrieved a lesion image (b), however, after joint optimization, the retrieved image is less directly related to the wrong label (c). In VQA-RAD, retrieval shifted from an unrelated medical condition (e) to an image depicting the condition of the query image.

domain adaptation. Our work differs by being the first to explore joint optimization optimized directly for downstream tasks, with minimal fine-tuning, as opposed to requiring large-scale pre-training. Our extensive novel analyses are the first to shed light on inconsistent predictions and how joint optimization on downstream tasks helps address them.

Multimodal Retrieval Augmentation in Medical Applications. Multimodal retrieval augmentation began with encoder-based architectures (Yuan et al., 2023), which integrated retrieved text and images with query modalities. With Large Vision-Language Models (LVLMs), He et al. (2024a) proposed retrieving relevant labeled examples during inference. Xia et al. (2024b) addressed LVLMs' misalignment with medical knowledge using retrieval-augmented generation to fetch similar medical reports, employing contrastive learning and context selection strategies. Xia et al. (2024a) proposed domain-aware retrieval for diverse medical data sources, while Wu et al. (2025) combined retrieval augmentation with knowledge graphs.

3 Method

Overview. In this section, we present our methodology, termed JOMED (Joint Optimization of Multimodal Retrieval for Enhanced Medical Diagnosis). Our task involves medical image classification and visual question answering. JOMED comprises two main components: a multimodal retriever and a reader. Given an input patient image and a diagnostic question, the multimodal retriever identifies relevant medical knowledge—images and their associated captions or hospital reports—that provide predictive information. The reader then analyzes the retrieved candidates along with the question to generate an answer. We jointly optimize the reader to better leverage the retrieved information and the multimodal retriever to enhance its ability to select informative candidates for the reader. Figure 3 illustrates our framework.

Unlike previous work (He et al., 2024a; Xia et al., 2024b,a) our method does not use medical pretraining, and instead we use readily available general LVLMs (Pixtral and Qwen2-vl) and a general retriever, jina-clip (Xiao et al., 2024). The selected LVLMs can process multiple images and texts, allowing us to incorporate both the retrieved image and the retrieved caption/report into the model context. We also compare our results to Med-Flamingo (Moor et al., 2023). Med-Flamingo is a medical LVLM which is also able to process multiple images, unlike other medical LVLM available at the time of conducting our experiments. We now elaborate on the details.

3.1 Reader Retrieval Augmentation Fine-Tuning

We first fine-tune the reader with retrieval augmentation. This step aims to achieve two main goals: improving the reader's performance on the dataset and teaching the reader to effectively utilize retrieved (image, text) pairs in context.

For an image-question tuple (i_d, q_d) where $d \in \{1, \ldots, D\}$ from dataset D, we retrieve K relevant tuples of images and texts (reports or captions) (i_k, t_k) where $k \in \{1, \ldots, K\}$ using our multimodal retriever module. To simplify notation, we denote $z_k = (i_k, t_k)$ as the k-th retrieved image-text pair and $q_d = (i_d, q_d)$ as the query pair for example d. These retrieved pairs are prepended to the query pair to create augmented inputs. The augmented input consists of the retrieved context followed by the target question-image pair: $z_k \circ q_d$ for $k \in \{1, \ldots, K\}$. We use supervised fine-tuning on the augmented input, minimizing the negative log-likelihood to predict the answer a_d:

$$\mathcal{L}(\theta) = -\sum_{d=1}^{D}\sum_{k=1}^{K} \log p_\theta(a_d \mid z_k \circ q_d)$$

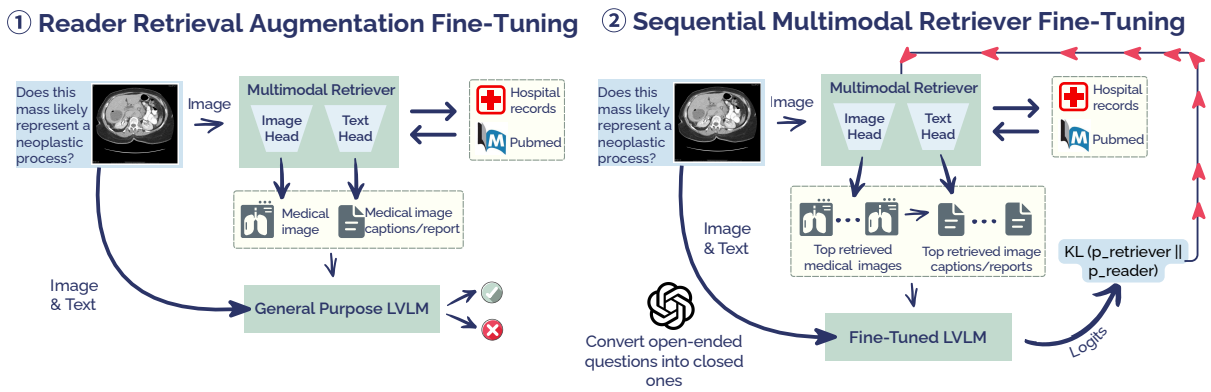

Figure 3: The two-phase JOMED training. The LVLM is first trained with a frozen retriever on augmented prompts containing retrieved image–caption/report pairs. With the LVLM frozen, the dual-head multimodal retriever is optimized using a KL divergence computed over a selected subset of the model's logits. An LLM converts open-ended questions into closed form (during training only), which improves results.

where θ represents the parameters of the LVLM, and $\circ$ denotes the concatenation operation.

3.2 Sequential Multimodal Retriever Fine-Tuning

Our dual-head multimodal retriever has a text-based head for retrieving relevant (image, text) pairs based on textual similarity to the query, and an image-based head for retrieving pairs based on visual similarity. To train this retriever, we adopt a sequential training strategy, first optimizing the text retrieval head followed by the image retrieval head. During this process, the reader model remains frozen, focusing solely on improving the retriever's embedding space. Similarly to Izacard et al. (2023) we minimize the KL divergence between the LVLM's posterior distribution over retrieved pairs and the retriever's distribution, however we do so directly on downstream tasks with only lightweight fine-tuning in the data-efficient regime without pre-training, and we customize the loss and the training recipe for open-ended questions leading to empirical gains.

Let $z_k = (\mathbf{i}_k, \mathbf{t}_k)$ denote the k-th retrieved image-text pair and $q = (\mathbf{i}, \mathbf{q})$ denote the query pair. The posterior distribution reflects the model's confidence in predicting the correct answer $\mathbf{a}$ given a retrieved pair z_k and query q:

$$p_k \propto p_{\text{LVLM}}(\mathbf{a} \mid z_k \circ q),$$

where $p_{\text{LVLM}}(\mathbf{a} \mid z_k \circ q)$ is the probability assigned by the LVLM to $\mathbf{a}$, and $\circ$ denotes prepending the retrieved chunk to the query.

We sharpen the LVLM's output distribution by restricting it to only the relevant class tokens from the benchmark's possible answers. Specifically, we extract the model's logits, select only the relevant tokens for each class, and apply a softmax to obtain a more discriminative distribution. We denote this class-restricted distribution as p_{LVLM_C} where $C = \{c_1, c_2, c_3, \ldots\}$ is the set of chosen class tokens. The normalized posterior p_k is formulated as:

$$p_k = \frac{\exp(\log p_{\text{LVLM}_C}(\mathbf{a} \mid z_k \circ q))}{\sum_{i=1}^{K} \exp(\log p_{\text{LVLM}_C}(\mathbf{a} \mid z_i \circ q))},$$

where K is the total number of retrieved pairs. The retriever's distribution is defined by:

$$p_{\text{RETR}}(z \mid q) = \frac{\exp(s(z,q)/\tau)}{\sum_{k=1}^{K} \exp(s(z_k,q)/\tau)},$$

where $s(z,q)$ is the similarity score between the pair and the query, and temperature τ controls distribution sharpness. For the text retrieval head, the score is computed using the dot product between the index caption/report embeddings and the query image embedding, while for the image retrieval head, we use the index image embeddings.

For our loss we compute the KL-divergence between retriever and reader $\text{KL}(p_{\text{LVLM}_C} \| p_{\text{RETR}})$:

$$\sum_{k=1}^{K} p_{\text{LVLM}_C}(z_k) \log \left(\frac{p_{\text{LVLM}_C}(z_k)}{p_{\text{RETR}}(z_k)} \right)$$

Finally, for visual question answering (VQA), we use the o3 model to convert open-ended questions into closed-ended ones (yes/no). See prompt

in Appendix E and analysis of approach contribution in Appendix D.1. We apply this only during training[2], without modifying questions at inference time. Reformatted questions will be made publicly available. This reformulation improves results, akin to restricting the LVLM output distribution.

3.3 Inference

Following Shi et al. (2023) for a given query pair $q = (\mathbf{i}, \mathbf{q})$, we retrieve the top-$K$ relevant chunks $z_k = (\mathbf{i}_k, \mathbf{t}_k)$ using the image as the query. Each chunk is prepended to the question, and the LVLM computes predictions for the augmented prompts in parallel. The final output probability is:

$$p_{\text{LVLM}}(a \mid q) = \sum_{k=1}^{K} p_{\text{LVLM}}(a \mid z_k \circ q) \cdot p_{\text{R}}(z_k \mid q),$$

where $p_{\text{R}}(z_k \mid q)$ is:

$$p_{\text{R}}(z_k \mid q) = \frac{\exp(s(q, z_k))}{\sum_{j=1}^{K} \exp(s(q, z_j))},$$

and $s(q, z_k)$ is the similarity score between the query and the retrieved chunk.

4 Experiments

4.1 Experimental Setup

Our datasets include real-world hospital datasets (BRSET (Nakayama et al., 2024) and VinDr-PCXR (Pham et al., 2022)) alongside a variety of classification and visual question answering datasets. We focus on a low-resource data-efficient setting (training sets ranging from 546 to 7007 samples in classification, 1790-19,700 in VQA). Medical image annotation is a resource-intensive task, demanding expert annotators whose availability is limited and expensive. Clinical AI also often has poor generalization across heterogeneous medical centers and patient populations, often requiring that models be trained for the specific context in which they are deployed (Casey, 2022). Such constraints often restrict researchers to datasets comprising only a few thousand samples for a given study. Retrieval augmentation is well-suited for this setting, as it is known to benefit low-data regimes the most by leveraging external knowledge.

We consider three types of classification tasks: binary, multi-class, and multi-label. These include BreastMNIST ("Breast"; breast ultrasound imaging, binary) (Al-Dhabyani et al., 2020), DermaMNIST ("Derma"; pigmented skin lesion images, multi-class) (Tschandl et al., 2018), RetinaMNIST ("Retina"; retinal fundus images with multi-class disease annotations) (Liu et al., 2022), and two challenging real-world multi-label datasets: VinDr-PCXR (a chest X-ray dataset with 15 labels) (Pham et al., 2022) and BRSET (an ophthalmology dataset with 14 labels) (Nakayama et al., 2024). For visual question answering, we use widely adopted benchmarks: VQA-RAD (Lau et al., 2018), SLAKE-English (Liu et al., 2021), and PathVQA (He et al., 2020). Full details are provided in Appendix A.1.

Retrieval augmentation is supported by an external index constructed from PubMed and medical records: PMC-OA (Lin et al., 2023), MIMIC-CXR (Johnson et al., 2019a), and ROCO (Rückert et al., 2024). Full descriptions are available in Appendix A.2. Our experiments leverage two backbone models—Pixtral (12B) (Agrawal et al., 2024) and Qwen2-vl (7B parameters) (Wang et al., 2024)—with jina-clip-v1 (Xiao et al., 2024) serving as the general-purpose retriever's visual head, embedding both the texts and images of the retrieval index. All runs were performed on a single L40s GPU. More details in Appendix B.

Baselines. We compare to several RAG baselines. RAD (He et al., 2024a), a retrieval-based approach for classification tasks, where the label of the most similar training image is used as context. We integrate RAD with both backbone models (Qwen2-VL and Pixtral). MMed-RAG (Xia et al., 2024a), a recent state-of-the-art multimodal RAG framework, using two backbones: LLaVA-Med and LLaVA, both paired with a medically pre-trained retriever (CLIP). We also adopt a standard fine-tuned RAG baseline in which the retriever is fixed and only the reader is fine-tuned, corresponding to the first phase of our approach. Finally, we compare our retriever loss with the base Perplexity Distillation Loss (PDist) (Izacard et al., 2023), integrated into JOMED in place of our customized retrieval loss. Full implementation details are in Appendix C. For comparison to medically pre-trained LVLMs, we include competitive baselines: BiomedGPT (Luo et al., 2023); three LLaVA-Med (Li et al., 2023) variants; two LVLM variants based on PMC CLIP (MedVInT-TE and MedVInT-TD) (Zhang et al., 2023b); and three InternVL(Chen et al., 2024) variants: MedDr, MedDr + RAD (He et al., 2024a), and GSCo (He et al.,

[2]Run once per question; the total cost is only a few dollars.

(a)

Backbone (Size)	Model	Breast		Derma		Retina		VinDr-PCXR		BRSET		Mean	
		ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
LLaVA (7B)	MMed-RAG	.85	.84	.75	.30	.63	.46	.55	.11	.42	.30	.64	.40
Qwen2-vl (7B)	Reader	.83	.77	.68	.27	.60	.44	.48	.08	.34	.23	.59	.36
	RAD	.84	.79	.43	.34	.55	.40	.57	.09	.40	.25	.56	.37
	FT RAG	.85	.82	.71	.42	.62	.48	.55	.09	.48	.27	.64	.42
	JOMED + PDist	.86	.83	.73	.47	.63	.49	.57	.09	.47	.29	.65	.43
	JOMED	**.87**	**.84**	**.76**	**.50**	**.65**	**.50**	**.57**	**.14**	**.49**	**.37**	**.67**	**.47**
	JOMED$_{oracle}$	.87	.84	.85	.64	.69	.51	.64	.14	.52	.41	.71	.51
Pixtral (12B)	Reader	.82	.77	.75	.52	.55	.44	.48	.08	.44	.33	.61	.43
	RAD	.85	.80	.75	.47	.56	.41	.48	.09	.47	.30	.62	.41
	FT RAG	.88	.85	.79	.60	.57	.47	.49	.09	.47	.33	.64	.47
	JOMED + PDist	.88	.84	.80	.62	.60	.47	.50	.09	.45	.35	.65	.47
	JOMED	**.90**	**.87**	**.80**	**.62**	**.60**	**.51**	**.56**	**.14**	**.51**	**.37**	**.67**	**.50**
	JOMED$_{oracle}$	.93	.90	.83	.67	.63	.54	.67	.15	.57	.41	.73	.53

(b)

Backbone (Size)	Model	VQA-RAD		SLAKE		PathVQA		Mean	
		Closed	Open	Closed	Open	Closed	Open	Closed	Open
LLaVA (7B)	MMed-RAG	.74	.39	.87	.81	.90	.31	.84	.50
Qwen2-vl (7B)	Reader	.73	.41	.84	.80	.87	.25	.81	.49
	FT RAG	.76	.45	.88	.81	.91	.33	.85	.53
	JOMED + PDist	.77	.45	.89	.81	.92	.32	.86	.53
	JOMED	**.79**	**.48**	**.90**	**.84**	**.93**	**.38**	**.87**	**.57**
	JOMED$_{oracle}$	.86	-	.92	-	.95	-	.91	-
Pixtral (12B)	Reader	.72	.38	.83	.80	.87	.26	.81	.48
	FT RAG	.74	.41	.88	.81	.88	.31	.83	.51
	JOMED + PDist	.75	.41	.89	.81	.89	.32	.84	.51
	JOMED	**.76**	**.45**	**.90**	**.84**	**.90**	**.36**	**.85**	**.55**
	JOMED$_{oracle}$	.88	-	.92	-	.95	-	.92	-

Table 1: Results for classification (a) and VQA (b) with non-medically pretrained LVLM backbones (see Tables 3, 4 for comparisons to with medical pre-training; MMed-RAG's retrieval component is medically pre-trained). joint optimization consistently leads to the best results. For VQA, we evaluate closed questions using the exact match metric and open questions using token recall. Using an oracle reranker (for classification and closed VQA; §4.2) demonstrates that even larger improvements are achievable with JOMED's top-retrieved images in some datasets.

	MMed-RAG	LLaVaMed variants	MedDr variants	GSCo	MedVInT variants
Medical Pre-Training Size	600K	600K	255K	255K	177K

Table 2: Medical pre-training size for LVLMs under extensive medical pretraining.

2024b). Details in Appendix C.

4.2 Results

Table 1a and Table 1b show the effectiveness of JOMED on five medical classification and three VQA benchmarks. We compare with methods of similar computational cost; all LVLMs were not extensively pre-trained on medical data, whereas the MMed-RAG retriever was. JOMED consistently outperforms MMed-RAG, RAD, the fine-tuned RAG baseline, JOMED + PDist loss, and the Reader baseline (LVLM without retrieval).

In addition, we evaluated our model in an oracle scenario. Instead of the final fusion of logits (Section 3.3), we applied an oracle that, given the model responses for different retrieved images, selects the correct answer if it exists.[3] Our findings in Tables 1a and 1b show that the correct answer exists in a large proportion of cases, yielding superior results. This shows that in these cases, JOMED is able to surface in its top retrieved images an image

[3]For open-ended questions we do not conduct this analysis as there is no simple boundary between correct/incorrect.

Model	Breast	Derma	VinDr	BRSET
MMed-RAG$_{\text{LLaVA-Med}}$	.89	.79	.11	.33
MedVInT-TE	.88	.78	-	-
MedVInT-TD	.90	.80	-	-
MedDr$_{\text{InternVL}}$	.72	-	.08	.08
MedDr + RAD$_{\text{InternVL}}$	.88	-	-	-
GSCo$_{\text{InternVL}}$	.93	-	.09	.33
JOMED $_{\text{Qwen2-vl}}$	.87	.76	.14	.37
JOMED$_{\text{Pixtral}}$	.90	.80	.14	.37

Table 3: Comparison to prior reported results of medical pre-trained models, for binary and multiclass classification (Breast and Derma; accuracy), and for multi-label classification (VinDr-PCXR and BRSET; F1).

Model	SLAKE		PathVQA	
	Closed	Open	Closed	Open
MMed-RAG$_{\text{LLaVA-Med}}$	.89	.84	.92	.39
BiomedGPT-B	.90	.85	.88	.28
LLaVA-Med$_{\text{LLaVA}}$	.83	.85	.91	.38
LLaVA-Med$_{\text{Vicuna}}$	.85	.83	.92	.39
LLaVA-Med$_{\text{BioMedCLIP}}$	.87	.87	.91	.39
JOMED$_{\text{Qwen2-vl}}$	.90	.84 (.84)	.93	.38 (.37)
JOMED$_{\text{Pixtral}}$	.90	.84 (.82)	.90	.36 (.35)

Table 4: Comparison to medical pre-trained models for visual question answering. We report the token recall metric, reported for LLaVA-Med variants and token F1 reported for BiomedGPT (in red).

which could lead to a correct prediction, however the information gets lost in the process of prediction fusion with other retrieved images. Simply taking the label with highest confidence or label with highest average confidence across retrieved images, underperforms. This motivates us to explore the use of the o3 multimodal reasoning model to detect this image. see Section 5.

Benchmarking versus medically pretrained LVLMs for classification. In Table 3, we include previously reported results for several leading medically pretrained LVLM methods (detailed in Section 4.1). JOMED demonstrates competitive performance compared to models that underwent extensive medical pretraining (pretraining sizes summarized in Table 2). JOMED with the Pixtral backbone matches or exceeds all models except GSCo on the Breast dataset. To our knowledge, no medically pretrained LVLMs have been evaluated on the Retina benchmark, and we are not aware of any prior LVLM results on our other classification benchmarks. For binary and multi-class classification, we report only accuracy (ACC), as prior works did not report F1 scores. Conversely, for multi-label classification, we report only F1 scores. We also evaluated a medical baseline, Med-Flamingo, which generally underperformed and showed limited benefit from retrieval augmentation. We attribute this to its relatively older LLaMA backbone (Touvron et al., 2023). Additionally, we experimented with using BiomedCLIP (Johnson et al., 2019b) as the retriever for our general-purpose LVLMs, which exhibited a similar trend to the non-medical retriever. Full results in Appendix D.3.

Benchmarking against pretrained LVLMs for visual question answering. In Table 4, we compare our approach with medically pretrained LVLM methods (detailed in Section 4.1) for VQA. JOMED achieves competitive performance relative to extensively pretrained models. For closed-question tasks, JOMED outperforms or matches existing models. For open-ended question answering, JOMED with the Qwen2-vl backbone matches LLaVA-Med$_{\text{Vicuna}}$ on the SLAKE dataset and LLaVA-Med$_{\text{LLaVA}}$ on the PathVQA dataset. Note that we do not report results on VQA-RAD, since our method was trained using an internal split of training and validation, whereas VQA-RAD provides only official training and test splits.

5 Analysis

JOMED boosts empirically challenging cases. We observed that the performance gap is substantially larger for predictions classified as inconsistent retrieval predictions (the proportion of inconsistent retrieval predictions is reported in Table 6). We define an inconsistent retrieval prediction as one in which the model predicts different labels for retrieved candidates given the same target query (see Figure 1). Specifically, if the model's predictions vary across retrieved candidates, the instance is categorized as an inconsistent retrieval prediction. Conversely, if the model consistently predicts the same label for all retrieved candidates, it is classified as a consistent retrieval prediction.

Both the consistent and inconsistent retrieval prediction sets are established after the initial reader training stage, and their performance is assessed following joint multimodal retriever fine-tuning. Detailed results are presented in Table 5. Overall, JOMED improves performance on inconsistent re-

Backbone		Model	Breast		Derma		Retina		VinDr		BRSET		VQA-RAD	SLAKE	**Mean**	**Mean**
			ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	Closed Acc	Closed Acc	**ACC**	**F1**
Qwen2-vl	Inconsistent	Reader	.40	.40	.50	.25	.35	.31	.41	.07	.33	.21	.43	.70	.40	.25
		RAG	.40	.40	.52	.36	.40	.28	.46	.07	.49	.25	.33	.84	.45	.27
		JOMED	**.80**	**.80**	**.62**	**.46**	**.52**	**.29**	**.48**	**.16**	**.50**	**.36**	**.47**	**.86**	**.58**	**.41**
	Consistent	Reader	.84	.77	.83	.29	.63	.45	.56	.09	.34	.22	.77	.85	.64	.36
		RAG	.86	.82	.90	.49	**.66**	.50	**.66**	**.10**	**.43**	.23	**.82**	.89	**.70**	.43
		JOMED	**.88**	**.84**	**.91**	**.50**	**.66**	**.55**	**.66**	**.10**	.41	**.37**	**.82**	**.91**	**.70**	**.47**
Pixtral	Inconsistent	Reader	.60	.59	.45	.33	.40	.32	.38	.07	.46	.28	.66	.64	.46	.32
		RAG	.72	.71	.44	.45	.35	.37	.38	.08	.39	.26	.72	.65	.46	.37
		JOMED	**.84**	**.84**	**.50**	**.50**	**.42**	**.46**	**.49**	**.13**	**.45**	**.35**	**.77**	**.68**	**.54**	**.46**
	Consistent	Reader	.86	.80	.80	.56	.56	.44	.66	.08	.61	.37	.74	.85	.70	.45
		RAG	**.91**	**.87**	.84	**.67**	**.62**	.50	**.71**	.08	**.64**	.38	**.75**	.91	.74	.50
		JOMED	**.91**	**.87**	**.85**	**.67**	**.62**	**.53**	**.71**	**.10**	**.64**	**.42**	**.75**	**.91**	**.75**	**.52**

Table 5: Analysis of joint optimization effect: for inconsistent retrieval predictions, we can see that applying joint optimization boosts performance significantly for all datasets. For consistent retrieval predictions, the results are slightly improved. Note that inconsistent retrievals are challenging to all models, even non-retrieval ones.

Model	Breast	Derma	Retina	VinDr	BRSET	VQARAD	SLAKE
Qwen2-vl	3%	51%	15%	50%	93%	12%	16%
Pixtral	16%	13%	16%	66%	70%	7%	8%

Table 6: Proportion inconsistent predictions cases, per dataset, for JOMED based on Qwen2-vl and Pixtral.

trieval predictions by +0.12 in accuracy and +0.13 in F1 score when using the Qwen2-vl backbone, and by +0.09 in both accuracy and F1 score with the Pixtral backbone, while also offering a slight improvement on the consistent. We also observed that inconsistent retrieval predictions are empirically more challenging for the reader, FT RAG and JOMED models, with a 10–20 point performance gap between the inconsistent and consistent.

Performance with o3 as reranker. Results for inconsistent predictions remain relatively low (Table 5), even after applying joint optimization (Table 5). However, in the oracle analysis reported above, we found that JOMED can sometimes retrieve images that lead to correct predictions, together with other images that lead to wrong predictions. We thus explore whether a state-of-the-art multimodal model, when used as an optional reranker, could close the gap toward oracle performance. Specifically, we investigated whether the o3 reasoning model (OpenAI, 2025) could effectively select the image from the retrieved set that offers the most predictive information, guiding the model toward correct classification. We provided o3 with an image to classify, along with four retrieved images+captions, and tasked it with identifying the one containing the most predictive information.

Our results show that o3's performance generally surpasses simply taking the image with highest confidence, but o3 is generally inferior compared to fusing logits in JOMED.We conclude that reranking in this setting remains a significant challenge even for a frontier model. Results in Appendix D.2.

Additional analyses and ablations. We evaluated the contribution of each component in our method in Appendix D.1. Additionally, we present several robustness ablations in Appendix D.4. We demonstrate robustness across different scenarios.

6 Conclusion

We demonstrated that joint optimization of general-purpose LVLMs and multimodal retrievers can achieve competitive medical diagnosis performance through lightweight, data-efficient fine-tuning, without medical pre-training. Future work may explore whether this approach generalizes as a cost-effective alternative to domain-specific pre-training when task-specific data and domain knowledge bases are available but large-scale pre-training is prohibitive. In addition, we identified and substantially mitigated inconsistent retrieval predictions, where different retrieved candidates lead to conflicting diagnoses. Our oracle analysis revealed a considerable performance gap between actual results and what was theoretically achievable using the retrieved images, providing a foundation for future research on closing this gap. Future work may also look into the prevalence and impact of inconsistent retrieval predictions more broadly.

Limitations

Sequential rather than end-to-end optimization. We acknowledge that, in our method, gradients do not flow simultaneously between the retriever and reader modules. Nevertheless, the reader leverages signals from the retriever during training to improve its performance, while the retriever, in turn, benefits from feedback provided by the reader to refine its retrieved content. This bidirectional interaction effectively induces a joint optimization effect within our framework.

Scope of evaluation. Our evaluation focuses on classification and visual question-answering tasks. We did not evaluate our method on report generation, which requires different capabilities (longer-form generation, clinical writing conventions) and would benefit from dedicated investigation. Additionally, while our method shows consistent improvements across diverse imaging modalities (ultrasound, fundus photography, X-ray, histopathology, dermatoscopy, CT) and tasks (visual question answering, malignant lesion classification, diabetic retinopathy severity grading), evaluation on additional specialized modalities (e.g., PET, Microscopy, OCT) and tasks (e.g., organ classification, retinal OCT disease classification) would further validate generalizability.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, and 1 others. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. 2020. Dataset of breast ultrasound images. *Data in brief*, 28:104863.

Ross Casey. 2022. [link].

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. 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*, pages 24185–24198.

Sunan He, Yuxiang Nie, Zhixuan Chen, Zhiyuan Cai, Hongmei Wang, Shu Yang, and Hao Chen. 2024a. Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. *CoRR*.

Sunan He, Yuxiang Nie, Hongmei Wang, Shu Yang, Yihui Wang, Zhiyuan Cai, Zhixuan Chen, Yingxue Xu, Luyang Luo, Huiling Xiang, and 1 others. 2024b. Gsco: Towards generalizable ai in medicine via generalist-specialist collaboration. *arXiv preprint arXiv:2404.15127*.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. 2023. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23369–23379.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43.

Daniel P Jeong, Pranav Mani, Saurabh Garg, Zachary C Lipton, and Michael Oberst. 2024. The limited impact of medical adaptation of large language and vision-language models. *arXiv preprint arXiv:2411.08870*.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019a. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019b. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Benjamin Lambert, Florence Forbes, Alan Tucholka, Senan Doyle, Harmonie Dehaene, and Michel Dojat. 2022. Trustworthy clinical ai solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis. *arXiv preprint arXiv:2210.03736*.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, and 1 others. 2024. Ra-dit: Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE.

Ruhan Liu, Xiangning Wang, Qiang Wu, Ling Dai, Xi Fang, Tao Yan, Jaemin Son, Shiqi Tang, Jiang Li, Zijian Gao, and 1 others. 2022. Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns*, 3(6).

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. 2023. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.

LF Nakayama, M Goncalves, L Zago Ribeiro, H Santos, D Ferraz, F Malerbi, and 1 others. 2024. A brazilian multilabel ophthalmological dataset (brset). 2023. *URL: https://physionet. org/content/brazilian-ophthalmological/1.0. 0/[accessed 2024-08-14]*.

OpenAI. 2025. Openai o3 and o4-mini system card.

H Hieu Pham, T Thanh Tran, and Ha Quy Nguyen. 2022. Vindr-pcxr: An open, large-scale pediatric chest x-ray dataset for interpretation of common thoracic diseases. *PhysioNet (version 1.0. 0)*, 10(2).

Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, and 1 others. 2024. Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1):688.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17.

Jiashuo Sun, Jihai Zhang, Yucheng Zhou, Zhaochen Su, Xiaoye Qu, and Yu Cheng. 2024. Surf: Teaching large vision-language models to selectively utilize retrieved information. In *EMNLP*.

Omkar Chakradhar Thawakar, Abdelrahman M Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Khan. 2024. Xraygpt: Chest radiographs summarization using large medical vision-language models. In *Proceedings of the 23rd workshop on biomedical natural language processing*, pages 440–448.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. 2018. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *arXiv preprint arXiv:2308.02463*.

Yinan Wu, Yuming Lu, Yan Zhou, Yifan Ding, Jingping Liu, and Tong Ruan. 2025. Mkgf: A multi-modal knowledge graph based rag framework to enhance lvlms for medical visual question answering. *Neurocomputing*, page 129999.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2024a. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. 2024b.

Rule: Reliable multimodal rag for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1081–1093.

Han Xiao, Georgios Mastrapas, and Bo Wang. 2024. Jina clip: Your clip model is also your text retriever. In *Multi-modal Foundation Model meets Embodied AI Workshop@ ICML2024*.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. 2023. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.

Zheng Yuan, Qiao Jin, Chuanqi Tan, Zhengyun Zhao, Hongyi Yuan, Fei Huang, and Songfang Huang. 2023. Ramm: Retrieval-augmented biomedical visual question answering with multi-modal pre-training. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 547–556.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, and 1 others. 2023a. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023b. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A Dataset

A.1 Dataset for Evaluation

Our datasets include real-world hospital datasets (BRSET (Nakayama et al., 2024) and VinDr-PCXR (Pham et al., 2022)) alongside a variety of classification and visual question answering datasets. We focus on a low-resource data-efficient setting (training sets ranging from 546 to 7007 samples in classification, 1790-19,700 in VQA). Medical image annotation is a resource-intensive task, demanding expert annotators whose availability is limited and expensive. Clinical AI also often has poor generalization across heterogeneous medical centers and patient populations, often requiring that models be trained for the specific context in which they are deployed (Casey, 2022). Such constraints often restrict researchers to datasets comprising only a few thousand samples for a given study. Retrieval augmentation is well-suited for this setting, as it is known to benefit low-data regimes the most by leveraging external knowledge to compensate for sparse training samples. For the classification tasks, we used the following datasets:

BreastMNIST (nickname: Breast) (Al-Dhabyani et al., 2020) is a binary classification dataset of breast ultrasound. Following the official split, we use 546 for training, 78 for validation and 156 for testing.

RetinaMNIST (nickname: Retina) (Liu et al., 2022) is a multi-label classification dataset of retina Fundus Camera. Following the official split, we use 1,080 for training, 120 for validation and 400 for testing.

DermaMNIST (nickname: Derma) (Tschandl et al., 2018) is a multi-class classification dataset of common pigmented skin lesions. Following the official split, we use 7,007 for training, 1,003 for validation and 2,005 for testing. The dataset contains clinical images from 7 different diagnostic categories: actinic keratoses, basal cell carcinoma, benign keratosis, dermatofibroma, melanoma, melanocytic nevi, and vascular lesions.

VinDr-PCXR (Pham et al., 2022) is a large, open pediatric chest X-ray dataset collected in Vietnam (2020–2021) containing 9,125 posteroanterior scans from patients under 10 years old. It provides both lesion-level bounding-box annotations for 36 findings and image-level labels for multi-label of 15 diagnoses, curated by experienced radiologists. We follow the official split, which includes 7,728 training and 1,397 test images, with an additional internal split of the training data into 85% for training and 15% for validation.

BRSET (Nakayama et al., 2024) is a Brazilian multilabel ophthalmological dataset comprising retinal fundus images annotated with 14 distinct pathological findings. The dataset, composed of 16,266 images, presents a challenging real-world scenario for multilabel classification. There is no publicly available split; therefore, we apply an internal split for train, validation, and test sets (70% train, 10% validation, and 20% test).

Breast, Derma, and Retina taken from the

(a)

Backbone (Size)	Model	Breast		Derma		Retina		VinDr-PCXR		BRSET		Mean	
		ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Qwen-vl (7B)	FT RAG	.85	.82	.71	.42	.62	.48	.55	.09	.48	.27	.64	.42
	Text Retriever Head	.86	.83	.74	.46	.62	.48	.56	.13	.48	.32	.65	.44
	JOMED	**.87**	**.84**	**.76**	**.50**	**.65**	**.50**	**.57**	**.14**	**.49**	**.37**	**.67**	**.47**
Pixtral (12B)	FT RAG	.88	.85	.79	.60	.57	.47	.49	.09	.47	.33	.64	.47
	Text Retriever Head	.89	.86	.79	.61	.58	.48	.55	.13	.50	.35	.66	.49
	JOMED	**.90**	**.87**	**.80**	**.62**	**.60**	**.51**	**.56**	**.14**	**.51**	**.37**	**.67**	**.50**

(b)

Backbone (Size)	Model	VQA-RAD		SLAKE		PathVQA		Mean	
		Closed	Open	Closed	Open	Closed	Open	Closed	Open
Qwen2-vl (7B)	FT RAG	.76	.45	.88	.81	.91	.33	.85	.53
	Text Retriever Head	.77	.46	.89	.82	**.93**	.35	.86	.54
	Closed Question Only	.78	.47	**.90**	.83	**.93**	.37	**.87**	.56
	JOMED	**.79**	**.48**	**.90**	**.84**	**.93**	**.38**	**.87**	**.57**
Pixtral (12B)	FT RAG	.74	.41	.88	.81	.88	.31	.83	.51
	Text Retriever Head	.76	.44	**.90**	**.83**	.89	.33	.85	.53
	Closed Question Only	.76	.45	**.90**	**.84**	.90	.36	.85	.55
	JOMED	**.78**	**.47**	**.90**	**.83**	**.93**	**.37**	**.87**	**.56**

Table 7: Ablation results for classification (a) and visual question answering (b). We demonstrate the necessity of each stage in our training paradigm. The first stage, FT RAG, trains only the reader. Subsequently, training the text retriever improves performance, and finally, completing the process with the image retriever (JOMED) yields the best results. We also evaluate training only on closed questions—without applying the o3 model to convert open questions into closed ones—and observe a gain in open-question performance in this setting

large-scale MNIST-like collection of standardized biomedical images, including 12 datasets for 2D and 6 datasets for 3D. We especially used MedMNIST+ (Yang et al., 2023), which is a higher resolution extension of the original MedMNIST. We used the highest resolution of 224 × 224.

For medical question answering, we have three well-known datasets:

VQA-RAD (Lau et al., 2018) is a clinician-annotated visual question answering dataset focused on radiology images (e.g., X-ray, CT, MRI). It pairs each image with both open-ended and yes/no questions. Following the official split, we use 1,753 questions for training and 453 for testing. In addition, we further partition the training set into 85% for training and 15% for validation.

PathVQA (He et al., 2020) is a pathology-focused medical VQA dataset built from textbooks and the PEIR digital library, comprising 4,289 images and 32,632 question–answer pairs spanning both open-ended and yes/no questions. We follow the official split provided by the authors: 19,700 for training, 6,260 for validation, and 6,720 for testing.

SLAKE (Liu et al., 2021) is a bilingual medical VQA dataset. We use the English subset, *SLAKE-English*, which comprises 642 radiology images and 7.03k English QA pairs. We follow the official train/validation/test split of 4.92k/1.05k/1.06k QA pairs, and include all question types, covering both open-ended and closed-ended formats.

A.2 Dataset for Index

For construction of the Index we used three large datasets of (image, text) pairs: PMC-OA (Lin et al., 2023), ROCO (Rückert et al., 2024) and MIMIC-CXR (Johnson et al., 2019a):

PMC-OA: is a large-scale dataset that contains 1.65M image-text pairs. The figures and captions from PubMed Central, 2,478,267 available papers are covered.

ROCO: is an image-caption dataset collected from PubMed. It filters out all the compound or non-radiological images, and consists of 81K samples.

MIMIC-CXR: is the largest chest X-ray dataset, containing 377,110 samples (image-report pairs). Each image is paired with a clinical report describing findings from doctors.

Backbone	Retrieval type	Breast		Derma		Retina		VQA-RAD	Mean	
		ACC	F1	ACC	F1	ACC	F1	ACC	ACC	F1
Flip a coin		.46	.30	.14	.10	.17	.16	.50	.35	.31
Pixtral	WO Q img	.75	.42	.66	.18	.47	.33	.67	.67	.49
	Reader only	.82	.66	.75	.53	.56	.44	.72	.74	.65
	WO Retrieval	.84	.65	.78	.55	.58	.48	.72	.74	.65
	Random Retrieval	.84	.69	.78	.56	.56	.45	.70	.74	.65
Qwen2-vl	WO Q img	.81	.52	.56	.20	.47	.20	.70	.68	.48
	Reader only	.82	.63	.68	.27	.61	.44	.74	.74	.59
	WO Retrieval	.86	.70	.70	.33	.62	.42	.76	.76	.62
	Random Retrieval	.85	.67	.73	.39	.63	.42	.77	.76	.62

Table 8: We evaluate our models in three scenarios: (1) the model is not given the query image and must rely solely on the retrieved context for prediction; (2) the model is not given any retrieval; and (3) the model is given a retrieved context that is randomly chosen and may mislead it. We use two baselines: a random classifier (uniformly selects a class) and a reader-only model (trained without retrieval). We evaluate on Breast, Derma, Retina, and the closed-question subset of VQA-RAD.

Backbone	Model	Breast		Derma		Retina	
		ACC	F1	ACC	F1	ACC	F1
Qwen2-vl	Top-1	.50	.60	.32	.25	.38	.32
	Top-1 logits	.60	.58	.43	.35	.37	.38
	o3	.77	.80	.42	.32	.35	.33
	o3 multi-image	.77	.80	.43	.33	.43	.35
	JOMED	.77	.80	.57	.43	.51	.47
Pixtral	Top-1	.72	.75	.46	.48	.29	.33
	Top-1 logits	.64	.63	.40	.40	.29	.32
	o3	.73	.73	.42	.37	.25	.30
	o3 multi-image	.73	.73	.43	.42	.34	.37
	JOMED	.76	.77	.45	.47	.34	.37

Table 9: Study of o3 as a reranker. We evaluate o3's ability to rerank the inconsistent predictions, selecting which of the retrieved (image, caption) pairs has the most predictive information for the query image. We compared the performance of choosing the highest similarity retrieved (image, caption) pair, choosing the highest confidence (top logits), using only the caption for reranking, and using our current strategy of fusing logits from all predictions. The metrics used are accuracy and F1 score.

B Implementation Details

B.1 Retrieval Augmentation

As described earlier, for each image–question pair, we retrieve r image–text pairs, with $r = 4$. We use the input image as the query and Jina-CLIP (Xiao et al., 2024) as the retriever head in our multimodal retriever. The index is constructed with FAISS (Johnson et al., 2019b) over MIMIC-CXR, PMC-OA, and ROCO, which are fully described in Section A.2 in Appendix A. To embed images, we use the Jina-CLIP visual head (the same model used for retrieval); to embed captions/reports, we use the Jina-CLIP text head. The index is stored in `float16` due to storage constraints. We also experimented with BiomedCLIP (Zhang et al., 2023a).

B.2 Reader Fine-Tuning

We use Pixtral (12B) (Agrawal et al., 2024) and Qwen2-vl (7B) (Wang et al., 2024) as base models. We train the models with and without retrieval augmentation to assess their effect. The retrieval-augmented prompt is described in Appendix E. Note, we also tried Med-Flamingo (9B) (Moor et al., 2023) as a base model.

We fine-tune Pixtral and Qwen2-vl using LLaMA-Factory (Zheng et al., 2024) on a single NVIDIA L40 (48 GB) GPU for 10 epochs. We use a learning rate of 2×10^{-5} and apply Low-Rank Adaptation (LoRA) (Hu et al., 2021) for parameter-efficient fine-tuning. To all models (including baseline models), we conduct a grid search over batch size (2, 4, 6) and whether to freeze the vision head, selecting the best configuration by validation performance. For Med-Flamingo, we fine-tune using our codebase on a single NVIDIA L40 (48 GB) GPU for 10 epochs. Following the Med-Flamingo paper, the language model and image encoder are frozen, and only the Gated Cross-Attention layers and the Perceiver Resampler are optimized for stable and efficient learning. We use a learning rate of 2×10^{-5}.

B.3 Multimodal Retriever Fine-Tuning

We train the retrieval model to fetch relevant context while keeping the reader frozen. We use Jina-CLIP's visual head as the base model for the multimodal retriever. The model is trained on a single NVIDIA L40 (48 GB) GPU for 100 epochs with a learning rate of 2×10^{-5}, freezing all layers except the last ten. We set the number of retrieved candidates per query to 50. In our experiments, retrieving more candidates further improved retriever performance.

B.4 Evaluation Metrics

Classification. We report Accuracy (ACC) and Macro F1:

$$\mathrm{ACC} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}[\hat{y}_n = y_n],$$

$$\mathrm{MacroF1} = \frac{1}{C}\sum_{c=1}^{C} \frac{2\,\mathrm{Prec}_c\,\mathrm{Rec}_c}{\mathrm{Prec}_c + \mathrm{Rec}_c},$$

where Prec_c and Rec_c are precision and recall computed per class c, C is the number of classes, and N is the number of examples.

Visual Question Answering (VQA). For open-ended questions, we compute token-level Recall and F1 between the predicted token set $\hat{A}$ and the reference token set A:

$$\mathrm{Prec} = \frac{|\hat{A} \cap A|}{|\hat{A}|}, \quad \mathrm{Rec} = \frac{|\hat{A} \cap A|}{|A|},$$

$$\mathrm{F1} = \frac{2\,\mathrm{Prec} \cdot \mathrm{Rec}}{\mathrm{Prec} + \mathrm{Rec}}.$$

For closed-ended questions, we use Exact Match.

C Baseline Methods

C.1 RAG baseline

We evaluate our method against several representative RAG-based and multimodal retrieval approaches to ensure a fair and comprehensive comparison.

RAD (He et al., 2024a) is a retrieval-based method designed for classification tasks. During inference, the most similar image from the training set is retrieved, and its supervised label is directly used as the prediction. We integrate RAD with both of our backbone models (Qwen2-VL and Pixtral) under the same data splits and retrieval settings.

MMed-RAG(Xia et al., 2024a) represents a recent state-of-the-art multimodal RAG approach in which the retriever and LVLM are trained independently. We follow the official GitHub implementation, first training the CLIP module and then performing DPO training using two backbone models: LLaVA-Med and LLaVA.

Fusion-in-Decoder (FiD) Pipeline (Izacard and Grave, 2020) serves as a standard RAG-style baseline. In this setting, the retriever is frozen, and only the reader is fine-tuned while processing retrieved image–caption pairs as contextual input.

Perplexity Distillation Loss (PDist) (Izacard et al., 2023). For each query and candidate document, we compute the reduction in perplexity (or equivalently, the increase in likelihood) of the correct output when the language model is conditioned on that document. From these likelihoods, we derive a "posterior" over documents (proportional to their contributions), and train the retriever to mimic that posterior via a KL divergence objective. In doing so, the retriever is guided to prioritize documents that most effectively support the language model's generation, thereby aligning retrieval with generation quality.

C.2 Medical Pre - trained baseline models

We benchmark against state-of-the-art medical large vision–language models (LVLMs) that underwent large-scale medical pre-training:

BiomedGPT. An open multimodal generative pre-trained transformer for biomedicine that aligns diverse biological/biomedical modalities with natural language; we include it as a strong domain baseline (Luo et al., 2023).

LLaVA-Med. A biomedical adaptation of LLaVA trained via curriculum (biomedical figure–caption alignment followed by instruction tuning); we evaluate three commonly used releases as separate baselines (Li et al., 2023).

MedVInT-TE and MedVInT-TD. Two implementations of Medical Visual Instruction Tuning from PMC-VQA (Zhang et al., 2023b). Both adopt a PMC-CLIP vision encoder (Lin et al., 2023). TE uses an encoder-style text pathway feeding a multimodal decoder, while TD concatenates text tokens with visual features to a decoder-only pathway; we report both as distinct baselines.

InternVL-based baselines: MedDr, MedDr+RAD, and GSCo. MedDr is a

Model	Breast		Derma		Retina		Vindr		BREST		VQARAD	SLAKE	PathVQA	Mean	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	ACC	ACC	ACC	F1
FT RAG$_{\text{Qwen2-vl}}$	.84	.81	.73	.38	.63	.45	.54	.09	.49	.27	.77	.89	**.92**	.72	.40
JOMED$_{\text{Qwen2-vl}}$	**.87**	**.82**	**.75**	**.40**	**.65**	**.50**	**.58**	**.15**	**.50**	**.39**	**.79**	**.91**	**.92**	**.74**	**.45**
FT RAG$_{\text{Pixtral}}$	.88	.84	.79	.57	.57	.46	.55	.09	.44	.31	.74	.87	.88	.72	.45
JOMED$_{\text{Pixtral}}$	**.91**	**.87**	**.80**	**.59**	**.59**	**.49**	**.56**	**.15**	**.50**	**.39**	**.77**	**.89**	**.89**	**.74**	**.50**

Table 10: Performance comparison of JOMED against FT RAG baseline across medical imaging classification and VQA tasks. Both Qwen2-VL and Pixtral backbones show consistent improvements with JOMED, particularly in F1 scores. Results demonstrate that BiomedCLIP-based retrieval enhances performance similarly to the general-purpose retriever. VQA results are reported for the closed-question subset using exact match accuracy.

Model	Breast		Derma		Retina		VQARAD
	ACC	F1	ACC	F1	ACC	F1	ACC
Reader only	**.82**	**.77**	.67	**.54**	.57	.40	.65
JOMED	.81	**.77**	**.68**	**.54**	**.60**	**.51**	**.69**

Table 11: Med-Flamingo comparison with and without JOMED retrieval augmentation across medical imaging tasks. JOMED achieves best performance on 4 out of 7 metrics, with notable improvements over Reader Only in Retina classification (ACC: 0.57→0.60, F1: 0.40→0.51) and VQA-RAD (0.65→0.69). However, the overall performance gains are more modest than those observed with other backbone architectures, indicating that Med-Flamingo may have limited capacity to leverage additional retrieval context.

generalist medical VLLM trained on large-scale instruction-style data curated from diagnosis-guided bootstrapping and medical image descriptions, covering multiple modalities and tasks (He et al., 2024a). **MedDr+RAD** augments MedDr at inference with Retrieval-Augmented Diagnosis (RAD): for a test image, similar cases are retrieved and summarized as contextual guidance in the prompt (He et al., 2024b). **GSCo** (Generalist–Specialist Collaboration) is a two-stage framework that (i) builds a generalist GFM (MedDr) and lightweight specialist models, and (ii) performs collaborative inference via Mixture-of-Expert Diagnosis (MoED; using specialist predictions as references) and RAD (using specialists to retrieve similar cases) (He et al., 2024b). GSCo evaluates across a large multi-dataset benchmark; we include GSCo as a strong InternVL-based baseline along with its MedDr and MedDr+RAD components (He et al., 2024b; Chen et al., 2024).

D Additional Analysis

D.1 Analysis of model components

We analyze the contribution of each part in the model in Table 7a and in Table 7b. The result of the first stage of training - the reader retrieval augmentation fine-tuning is the FT RAG then we check the contribution of training only the text retriever head and then finally is the contribution of training both the text and image retriever head as JOMED. We also explore the utility of converting the open question into a closed one in our retrieval loss. If we don't apply o3 model to convert the open question, we only trained on the closed one, which is referred to in Table 1b as Closed Question Only.

D.2 Performance with a State-of-the-Art LVLM Reranker

We evaluate whether a state-of-the-art large vision–language model (LVLM), used as an optional reranker, can improve performance on inconsistent-retrieval predictions. Specifically, we test the o3 reasoning model (OpenAI, 2025) as a reranker: given a query image to classify and four retrieved image–caption pairs, o3 is asked to select the single pair that is most informative for predicting the correct label. We compare this reranking to four alternatives: (i) using the top-1 retrieved candidate (no reranking), (ii) selecting the model's prediction with the highest confidence (maximum logit), (iii) using o3 with captions only (no images), and (iv) the JOMED aggregation strategy.

Our results show that o3 generally outperforms the maximum-logit baseline but remains inferior to JOMED's logit-fusion aggregation. We evaluated performance on the Breast, Derma, and Retina

datasets. For the remaining datasets—VinDr-PCXR, BRSET, and the visual question answering (VQA) datasets—we did not run this evaluation because the images and/or retrieved candidates come from restricted-access sources (e.g., MIMIC-CXR, VinDr-PCXR, and BRSET). See results in Table 9.

D.3 JOMED variants with medical pre-trained backbones

Although our work focuses on general-purpose LVLMs, we also evaluated a medical-specific baseline. We conducted the evaluation on four benchmarks—Breast, Derma, Retina, and VQA-RAD (closed-question subset). For the medical pre-trained baseline, we used Med-Flamingo and paired it with a medically pre-trained retriever (BiomedCLIP), using only one retrieval head (image retrieval without text retrieval). The performance gap between Med-Flamingo and our reader approach was small—0.01 in accuracy and 0.03 in F1. We hypothesize that this narrow gap may be due to Med-Flamingo's relatively older LLaMA backbone (Touvron et al., 2023). Results are presented in Table 11. We also tested our general-purpose backbone with a medically pre-trained retriever, which showed improvements over FT RAG, with a similar pattern to the general-purpose retriever. Full results are presented in Table 10.

D.4 Retrieval Robustness Analyses.

We strive for our model to be robust in cases where the retrieval is noisy. Meaning, that the intrinsic ability in the model parameters is not overly dependent on the retrieval content. To test this property without ground truth retrieval labels, we evaluate the model performance with random candidates instead of searching according to cosine similarity. Overall as shown in Table 8 in Appendix D, we observed that for all datasets the model performance is above the reader-only, showing the model succeeded more than a model trained without any retrieval.

We also test the ability of the model to still function without any retrieved context. We do this to check whether some intrinsic capability is still maintained in the model's reader without any retrieved information. This could also be helpful in cases where model users would like it to be versatile and support both retrieval-augmented prediction and non-retrieval-augmented prediction (e.g., in cases where no retrieved contexts were discovered). As shown in Table 8 in Appendix D, removing retrieval from JOMED resulted in performance comparable or slightly lower than the reader-only model.

Finally, we tested the model without the input image, relying only on the (image, report / caption) retrievals, to further assess the information held in retrieved data for predictions. In this scenario, we expected a substantial performance drop because the model no longer had direct knowledge of the patient's condition, aside from what was provided by retrieved candidates. As a baseline, we used a simple random guess (coin flip), representing a scenario in which retrieval was not used effectively. As shown in Table 8 in Appendix D, our model's performance exceeded this baseline by a large margin, with improvements of 0.32 in Accuracy, 0.28 in F1. These results indicate that the model learns from the retrieved data, yet it does not become overly dependent on it as we previously showed.

E Prompt Design

We detail the prompts used in our experiments and analyses. To convert open questions into closed ones with the o3 model, we use the prompt in Prompt 1. To assess the utility of o3 as a re-ranker of the most predictive image–caption pair (Section D.2, Appendix D), we provide both LVLM backbones (Pixtral and Qwen2-VL) with the task-specific prompt in Prompt 2.

We also include the training prompts for JOMED across datasets—Prompts 3, 5, 4, 7, 6—and the VQA prompt (Prompt 8).

```
    Convert the following open-ended question into a closed yes/no question based on the given answer
    .
The new question should be answerable with "{expected_answer}".

Original Question: {question}
Original Answer: {answer}
Expected New Answer: {expected_answer}

Please provide only the new closed question without any additional text or explanation.
```

Listing 1: Prompt used for o3 to convert open question into closed one

```
    Task:

    You are an expert assistant selecting the most informative reference for medical image
     classification.

    Description:
    A patient's [MRI/CT/X-ray] image needs to be classified into one of the following categories: [
     List of possible labels].
    You are given four candidate reference pairs, each consisting of a caption and its associated
     image, drawn from PubMed literature.
    Your goal is to determine which single candidate provides the most useful information to help an
     AI model correctly classify the provided image.

    Input Information:
    - Medical image: The patient's MRI/CT/X-ray image is provided here.
    - Classification task: Classify the image into one of the following categories: [List of possible
      labels].
    - Candidates:
        - Candidate 0:
            - Caption: [Caption 0]
            - Image: [Image 0]
        - Candidate 1:
            - Caption: [Caption 1]
            - Image: [Image 1]
        - Candidate 2:
            - Caption: [Caption 2]
            - Image: [Image 2]
        - Candidate 3:
            - Caption: [Caption 3]
            - Image: [Image 3]

    Instructions:
    1. For each candidate, carefully assess how informative its caption and image are for the current
      classification task.
    2. Specifically, evaluate whether the candidate describes imaging features, findings, or clinical
      context relevant to distinguishing among the listed categories.
    3. Compare all candidates and select the one that gives the clearest, most discriminative
     information to aid the classification.
    4. Output only the number of the selected candidate in the following format:
       Caption Number: [your selected number]
```

Listing 2: Prompt used for o3

```
    <retrieved image>background:<retrieved
text>\n\n<query image>Does this breast
ultrasound image show signs of cancer?
```

Listing 3: Prompt used by JOMED for Breast classification

```
    <retrieved image>background:<retrieved
text>\n<query image>what is the severity of
diabetic retinopath?
```

Listing 4: Prompt used by JOMED for Retina classification

```
        <retrieved image>background:<retrieved text>\n\nProvide answer according to the labels:\n"
          "0 - 'Actinic keratoses and intraepithelial carcinoma'\n"
          "1 - 'Basal cell carcinoma'\n"
          "2 - 'Benign keratosis-like lesions'\n"
          "3 - 'Dermatofibroma'\n"
          "4 - 'Melanoma'\n"
          "5 - 'Melanocytic nevi'\n"
          "6 - 'Vascular lesions'\n\n"<query image>What type of skin lesion does the patient have?
```

Listing 5: Prompt used by JOMED for Derma classification

```
       <retrieved image>background:<retrieved text> Provide answer according to the labels:\n
       0 - 'No finding'\n
       1 - 'Bronchitis'\n
       2 - 'Brocho-pneumonia'\n
       3 - 'Other disease'\n
       4 - 'Bronchiolitis'\n
       5 - 'Situs inversus'\n
       6 - 'Pneumonia'\n
       7 - 'Pleuro-pneumonia'\n
       8 - 'Diagphramatic hernia'\n
       9 - 'Tuberculosis'\n
       10 - 'Congenital emphysema'\n
       11 - 'CPAM'\n
       12 - 'Hyaline membrane disease'\n
       13 - 'Mediastinal tumor'\n
       14 - 'Lung tumor'\n\n
       \n\n<image>Question: Look at this X-ray scan and select all abnormalities you see from the
     given labels Answer:
```

Listing 6: Prompt used by JOMED for Vindr-PCXR classification

```
<retrieved image>background:<retrieved text> Provide answer according to the labels:\n
0 - 'no findings'\n
1 - 'diabetic_retinopathy'\n
2 - 'macular_edema'\n
3 - 'scar'\n
4 - 'nevus'\n
5 - 'amd'\n
6 - 'vascular_occlusion'\n
7 - 'hypertensive_retinopathy'\n
8 - 'drusens'\n
9 - 'hemorrhage'\n
10 - 'retinal_detachment'\n
11 - 'myopic_fundus'\n
12 - 'increased_cup_disc'\n
13 - 'other'\n\n
\n\n<image>Question: Look at retinal fundus image and select all abnormalities you see from the given
      labels
```

Listing 7: Prompt used by JOMED for BRSET classification

```
<retrieved image>background:<retrieved text>\n<query image><query question>?
```

Listing 8: Prompt used by JOMED for visual question answering tasks