Retina-RAG: Retrieval-Augmented Vision–Language Modeling for Joint Retinal Diagnosis and Clinical Report Generation

Abdelrahman Zaian, Sheethal Bhat, and Andreas Maier

Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
{abdelrahman.zaian, sheethal.bhat, andreas.maier}@fau.de

Abstract. Diabetic Retinopathy (DR) is a leading cause of preventable blindness among working-age adults worldwide, yet most automated screening systems are limited to image-level classification and lack clinically structured reporting. We propose Retina-RAG, a low-cost modular framework that jointly performs DR severity grading, macular edema (ME) detection, and report generation. The architecture decouples a high-performance retinal classifier and a parameter-efficient vision–language model (Qwen2.5-VL-7B-Instruct) adapted via Low-Rank Adaptation (LoRA), enabling flexible component integration. A retrieval-augmented generation (RAG) module injects curated ophthalmic knowledge together with structured classifier outputs at inference time to improve diagnostic consistency and reduce hallucinations. Retina-RAG achieves an F1-score of 0.731 for DR grading and 0.948 for ME detection, substantially outperforming zero-shot Qwen (0.096, 0.732) and fine-tuned MMed-RAG (0.541, 0.641) on a retinal disease detection dataset with captions. For report generation, Retina-RAG attains ROUGE-L 0.438 and SBERT similarity 0.884, exceeding all baselines. The full framework operates on a single consumer-grade GPU, demonstrating that clinically structured retinal AI can be achieved with modest computational resources. Source code: https://github.com/zaian1st/RETINARAG

Keywords: Diabetic Retinopathy, Retrieval-Augmented Generation, LoRA Vision-Language Models, Medical Report Generation

1 Introduction

Diabetic retinopathy (DR) is a leading cause of preventable blindness among working-age adults worldwide [1]. The Global Burden of Disease Study 2021 projects DR-attributable vision impairment to rise by 55.6% to 160.50 million cases by 2045 [2], driven by urbanisation, ageing, and the expanding diabetes epidemic in low- and middle-income countries [1]. Yet access to trained ophthalmologists remains severely limited in many high-prevalence regions [2], creating an urgent need for scalable, automated screening that is accurate and clinically interpretable.

Although deep learning achieves strong DR grading performance, most approaches remain limited to image-level classification without clinically interpretable reports. CNN architectures such as EfficientNet [3] and ResNet [4] grade competitively even under class imbalance [5]. This limitation persists commercially: LumineticsCore (formerly IDx-DR), the first FDA-cleared autonomous DR-detection system in U.S. primary care, outputs only a binary referral decision without any structured description of retinal findings [6], limiting translational value in ophthalmic workflows.

Vision–Language models (VLMs) such as LLaVA-Med extend instruction tuning to medical imaging [7], but both general-purpose and medically adapted VLMs frequently produce factually incorrect outputs [8], with consistent hallucination patterns across modalities [9]; adapting large VLMs also demands substantial compute [10], motivating parameter-efficient alternatives. Retrieval-Augmented Generation (RAG) improves factual reliability by conditioning generation on retrieved knowledge [11]: RULE aligns retrieval via preference learning [12], and MMed-RAG adds domain-aware retrieval [13]. However, existing RAG frameworks are not designed for DR screening, which combines imbalanced multi-class severity grading with binary Macular Edema (ME) detection.

To address this gap, we introduce **Retina-RAG**, a modular framework for joint DR severity grading, ME detection, and structured report generation. Retina-RAG, as seen in Fig. 1 separates a high-performance retinal classifier from a parameter-efficient VL backbone and integrates classifier-guided retrieval prompting, injecting structured diagnostic predictions alongside curated ophthalmic knowledge at inference. This design aligns language generation with task-specific classification outputs, improving diagnostic consistency while avoiding computationally expensive end-to-end (e2e) multimodal training.

Main contributions: We 1) propose Retina-RAG, a diagnostic-aware framework that separates DR severity grading and ME detection from report generation, enabling modular optimization while preserving clinical task alignment. 2) introduce a classifier-guided retrieval prompting mechanism that conditions language generation on structured diagnostic predictions and curated ophthalmic knowledge, enforcing consistency between classification and reporting without costly e2e multimodal training. 3) provide comprehensive evaluation on a retinal disease and caption dataset [14], demonstrating improved grading (F1: 0.731), ME detection (F1: 0.948), and report quality (ROUGE-L 0.438; SBERT 0.884) using Retina-RAG in comparison to zero-shot (ZS) and medical RAG baselines.

2 Materials and methods

2.1 Dataset details and preparation

We evaluate Retina-RAG on the Retinal Disease Detection (RDD) dataset [14]: 2,254 colour fundus images (1,577 train / 339 val / 338 test) at ≈2,000×1,333 px,

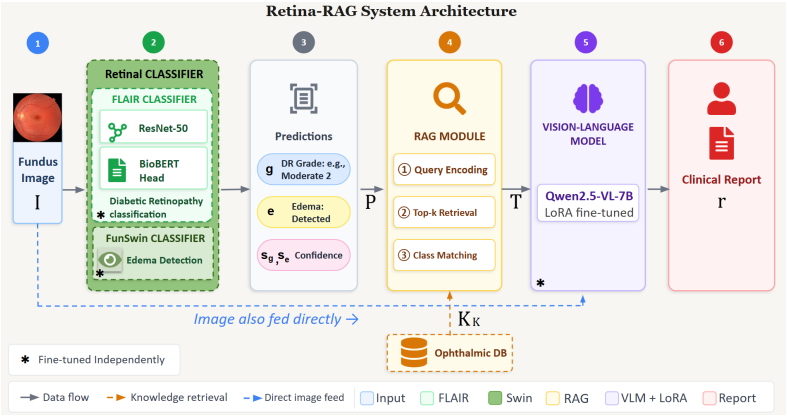

Fig. 1. Overview of Retina-RAG for both training and inference. We use a dual-branch retinal classifier to process the fundus image. The classifier predicts either DR severity and ME, producing structured text outputs $\mathcal{P}$. These predictions are serialized to query an ophthalmic knowledge base, retrieving k task-relevant snippets $\mathcal{K}_k$. The fundus image, classifier outputs, and retrieved knowledge are composed into a classifier-guided prompt $\mathcal{T}$ for a LoRA-adapted VLM (Qwen2.5-VL), which generates the final clinical report.

each with a five-class DR grade and binary ME label. Grades are severely imbalanced (No DR 52.6%, Moderate 22.7%, Mild 13.0%, Severe 7.4%, Proliferative 4.3%; ME 19.6% positive). MESSIDOR [15] contributes 1,200 images at 2,240×1,488 used solely as supplementary ME-branch training data, with no patient overlap. Class imbalance is addressed via inverse-frequency weighting. Images are resized to 512×512 (DR), 224×224 (ME), and 448×448 (VLM) and normalised with ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$); classifiers use bilinear interpolation and the VLM uses Lanczos resampling. ME inputs additionally apply morphological top-hat/black-hat transforms (15×15 elliptical kernel) on the green channel for lesion contrast.

2.2 Methodology

Retina-RAG is a modular three-stage pipeline (Fig. 1) comprising a dual-branch retinal classifier, a classifier-guided RAG module, and a LoRA-adapted VLM. The classifier and VLM are trained independently, enabling low-resource adaptation and component replacement. Given a fundus image $\mathcal{I}$, the classifier produces a DR grade $g \in \{0, \ldots, 4\}$ and ME label $e \in \{0, 1\}$ with confidences s_g, s_e; the structured output $\mathcal{P} = \{g, e, s_g, s_e\}$ then drives retrieval of k knowledge snippets $\mathcal{K}_k$, and the VLM generates the final report from $\mathcal{T} = \{\mathcal{P}, \mathcal{K}_k\}$.

(i) Retinal classifier. The classifier consists of two independently fine-tuned branches selected for their performance on DR and ME, respectively. For DR

grading we adopt FLAIR [16], an ophthalmology foundation model pretrained on 288,307 fundus images across 101 ophthalmic labels that combines a ResNet-50 encoder with a BioClinicalBERT text head via contrastive pretraining; we adapt it for five-class grading ($g \in \{0, \ldots, 4\}$: No DR, Mild, Moderate, Severe, Proliferative) by linear probing, freezing the backbone and training a single linear head with cross-entropy loss. Binary ME detection uses an ImageNet-pretrained FunSwin Transformer [30] fine-tuned with binary cross-entropy loss. Decoupling the two branches preserves each backbone's domain-specific representations and allows per-task optimization; at inference both branches run in parallel and their outputs are serialised into $\mathcal{P} = \{g, e, s_g, s_e\}$, which is passed downstream to both the RAG module and the VLM prompt.

(ii) Classifier-guided RAG module. The RAG module grounds report generation in structured classifier outputs and retrieved ophthalmic knowledge, reducing hallucination and enforcing consistency between predicted grades and generated text [12]. Its offline knowledge base $\mathrm{KB} = \{(\mathbf{d}_i, c_i)\}_{i=1}^{M}$ comprises curated DR clinical descriptions grounded in the ICDR scale [17], AAO Preferred Practice Patterns [18], and Joslin Diabetes Center guidelines, where each entry pairs a snippet t_i (describing pathological features, clinical significance, and management for a given DR grade or ME status) with its class label $c_i \in \{0, \ldots, 4\} \cup \{0, 1\}$. Snippets are stored in a hybrid index combining a dense FAISS [19] index over `MedEmbed-small-v0.1` embeddings ($\mathbf{d}_i = \phi(t_i) \in \mathbb{R}^d$) for semantic retrieval with a BM25 (Okapi) sparse index for keyword matching, with source documents split into chunks of 500 characters (50-character overlap) across the five DR grades and both ME states. At inference, $\mathcal{P}$ is serialized into a query (e.g., *"Moderate DR, ME detected, confidence 0.87"*) and encoded as $\mathbf{q} = \phi(q_{\mathrm{str}})$, after which retrieval is restricted to the class-matched subset $\mathrm{KB}_g = \{(\mathbf{d}_i, c_i) : c_i = g \text{ or } c_i = e\}$ a hard constraint ensuring retrieved snippets match the predicted grade g or ME status e and preventing clinically inconsistent context from non-matching grades. Within KB_g, the dense and BM25 indices each return their top 20 candidates, which are merged via Reciprocal Rank Fusion (with a soft boost for entries surfaced by both retrievers) into 20–30 unique candidates, re-scored by a CrossEncoder reranker to the top 6, of which the top-k (k=3) snippets $\mathcal{K}_k$ are retained; this dense–sparse hybrid with reranking improves precision over single-retriever cosine search while class matching preserves diagnostic consistency. Finally, the fundus image $\mathcal{I}$, classifier outputs $\mathcal{P}$, and retrieved snippets $\mathcal{K}_k$ are composed into a structured prompt $\mathcal{T} = \texttt{Prompt}(\mathcal{I}, \mathcal{P}, \mathcal{K}_k)$. The prompt casts the VLM as an expert ophthalmologist, supplies $\mathcal{P}$ as the primary diagnostic anchor and $\mathcal{K}_k$ as supporting context, and instructs it to examine the image first, treat the automated reading as one data point, and override it when the visual evidence conflicts (fine-tuned VLM only), then emit a fixed-format report (`RETINOPATHY_GRADE`:[0–4]; `MACULAR_EDEMA_RISK`:[0–1]; detailed findings) covering the retinal vessels, macula, and peripheral retina. This image-first override is enabled only with the

fine-tuned VLM, which has learned to weigh visual evidence against the anchor, whereas a frozen VLM treats the anchor as authoritative.

The condensed prompt template is: *"You are an expert ophthalmologist analyzing retinal fundus images. The classifier predicts: {$\mathcal{P}$}. Relevant clinical context: {$\mathcal{K}_k$}. Classify the image and generate a concise diagnostic report. Critical instructions: (1) examine the image first and identify all visible pathological features; (2) form your clinical assessment based on what you observe in the fundus; (3) consider any automated analysis as one data point, not ground truth; (4) if your observations conflict with automated readings, trust your visual analysis (fine-tuned VLM only). Systematically evaluate the retinal vessels (caliber, tortuosity, arteriovenous nicking), macula (foveal reflex, hard exudates, edema, hemorrhages), and peripheral retina (microaneurysms, cotton-wool spots, IRMA, neovascularization). Provide the assessment in the exact format:* `RETINOPATHY_GRADE`*:[0–4];* `MACULAR_EDEMA_RISK`*:[0–1]; detailed clinical findings (clinical assessment, macular assessment, findings, diagnosis)."*

(iii) LoRA-adapted VLM. Report generation uses Qwen2.5-VL-7B-Instruct [20], chosen for its strong instruction-following under resource constraints [10]. We fine-tune it with Unsloth's `FastVisionModel` [21] and LoRA [22] applied to attention and MLP projections in both the vision and language blocks, where the adapted weight update is

$$W' = W + \tfrac{\alpha}{r} BA, \tag{1}$$

with W frozen and $A \in \mathbb{R}^{r\times d}$, $B \in \mathbb{R}^{d\times r}$ trainable ($r \ll d$). Training uses supervised fine-tuning [31, 21] with a causal-language-modelling loss computed only over the assistant response tokens.

3 Experiments

Evaluation protocol: Classification under class imbalance is evaluated with weighted-averaged F1, precision, and recall (AUC macro-averaged across DR categories); accuracy is omitted as misleading under skew. Report quality uses BLEU-4 [32], ROUGE-1/ROUGE-L [33], and SBERT [34]/ClinicalBERT [35] similarity. These capture lexical/semantic proximity but not clinical actionability; expert factuality assessment is left to future work.

Implementation details: FLAIR is fine-tuned for 50 epochs using Adam (LR 1×10^{-4} for encoder–decoder, 1×10^{-5} for visual extractor, weight decay 1×10^{-5}) [22], while FunSwin uses AdamW (LR 1×10^{-4}) [30]. Qwen2.5-VL-7B-Instruct is adapted via LoRA (r=64, α=128) applied to attention and MLP layers in both vision and language branches; the 4-bit base model [24] is frozen and only LoRA parameters are updated. Training employs 8-bit AdamW with cosine decay (warmup 0.15), weight decay 0.01, batch size 2 with gradient accumulation (×4), and gradient clipping 0.5; loss is computed only over assistant-response

Table 1. Comparison of Retina-RAG against SOTA methods on the RDD test set: (a) classification (Prec./Rec./F1 weighted-averaged, AUC macro-averaged); (b) report-generation quality. MMed-RAG and FFA-IR are fine-tuned on the same RDD training split; results are mean of 3 runs.

(a) Classification

Method	DR Severity (5-class)				ME (Binary)			
	Prec.	**Rec.**	**F1**	**AUC**	**Prec.**	**Rec.**	**F1**	**AUC**
ZS Qwen [26]	0.076	0.204	0.096	0.511	0.743	0.722	0.732	0.598
MMed-RAG [13]	0.679	0.450	0.541	0.603	0.682	0.621	0.641	0.494
FFA-IR [27]	0.668	0.650	0.660	**0.837**	0.942	0.945	0.947	**0.934**
Retina-RAG (Ours)	**0.778**	**0.795**	**0.731**	0.774	**0.949**	**0.947**	**0.948**	0.897

(b) Report generation (NLP metrics)

Method	BLEU-4	ROUGE-1	ROUGE-L	SBERT	Clin-BERT
ZS Qwen [26]	0.014	0.175	0.163	0.764	0.921
MMed-RAG [13]	0.082	0.365	0.353	0.786	0.923
FFA-IR [27]	**0.235**	**0.529**	0.383	0.869	**0.982**
Retina-RAG (Ours)	0.194	0.461	**0.438**	**0.884**	0.975

tokens (SFT). Minority DR grades are upsampled with $\pm 15^\circ$ rotation and brightness augmentation. The RAG module fuses dense (`MedEmbed-small-v0.1`/FAISS) and sparse (BM25 Okapi) retrieval via RRF, reranks with a CrossEncoder, and passes the top k=3 snippets to the VLM. Experiments run on a single RTX 1650 (4GB), and results report the mean of three seeds.

Baselines: We compare Retina-RAG against (1) ZS Qwen (unmodified Qwen2.5-VL-7B-Instruct with a generic ophthalmology instruction, no fine-tuning or retrieval), (2) MMed-RAG [13], and (3) FFA-IR [27]. MMed-RAG and FFA-IR are fine-tuned on the same RDD training split as Retina-RAG, ensuring a fair in-domain comparison. We do not compare against keyword-driven captioning models on DeepEyeNet [36], which condition on expert-provided keywords and target free-form captioning rather than joint ICDR grading and ME detection, making a direct comparison confounded by task and label structure.

3.1 Results

Table 1a reports classification performance where Retina-RAG's DR grade and ME risk are extracted from the VLM's generated report (the `RETINOPATHY_GRADE` and `MACULAR_EDEMA_RISK` output fields), since the fine-tuned VLM is instructed to weigh its own visual analysis against the classifier anchor $\mathcal{P}$ (Sec. 2.2); this differs from the raw classifier output shown in Table 2, and explains why Retina-RAG's reported F1 (0.731) can diverge slightly from "Classifier + FT VLM, FLAIR" (0.698). Retina-RAG attains the best classification, with weighted-F1 0.731 (DR) and 0.948 (ME), substantially above ZS Qwen (0.096) and fine-tuned MMed-RAG (0.541) and improving on FFA-IR (0.660) for grading, while

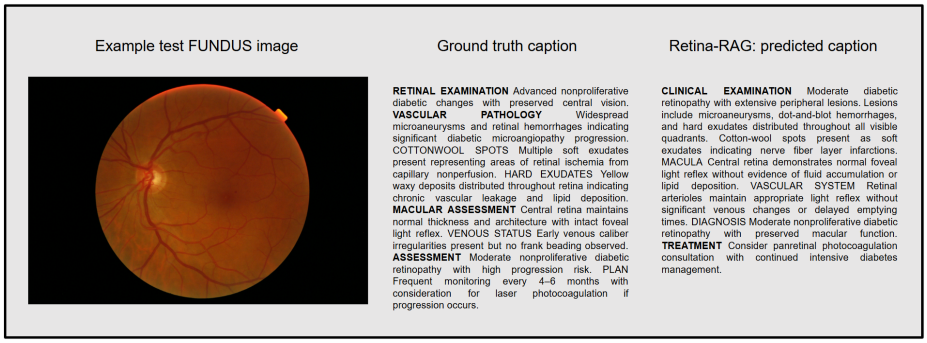

Fig. 2. Qualitative example from RDD test dataset (Image ID 00804) with groundtruth and Retina-RAG predicted report. Image indicates DR Grade 2 (Moderate); ME Risk 0 (No Risk), which is reflected in the predicted report.

competitive in AUC (0.774 vs. 0.837). For ME it leads all methods (0.948), underscoring the value of domain adaptation and retrieval grounding.

Table 1b reports report-generation quality, our primary evidence for the framework as a whole. Retina-RAG reaches ROUGE-L 0.438 and SBERT 0.884, well above MMed-RAG (0.353/0.786) and ZS Qwen (0.163/0.764), with ClinicalBERT 0.975 confirming clinical coherence. FFA-IR is limited to short reports ($\leq$60 words), whereas Retina-RAG generates long-form ($\sim$1,000-word) reports at comparable quality.

Diagnostic anchoring and error propagation: To test whether the VLM follows the classifier anchor, we extract the DR grade from each generated report and compare it to the classifier prediction. With a frozen VLM (classifier + ZS Qwen), the model diverges sharply: Grade-0 accuracy falls from 0.94 at the classifier to 0.29, with 112 of 178 true Grade-0 cases reallocated to Grade 1, indicating the anchor is treated as noise. After fine-tuning, alignment is largely restored (Grade-0 0.94, Grade 2 0.80–0.81, Grade 4 0.67$\rightarrow$0.73); the residual gap at Grade 3 (0.44 vs. 0.64) reflects the visual overlap between severe nonproliferative and proliferative disease, where classifier errors are most likely to propagate. The dominant failure mode is therefore one-grade underestimation at advanced stages rather than arbitrary hallucination. Isolating retrieval, the classifier + FT VLM and full Retina-RAG configurations differ only in whether the RAG module is active (Table 2b): retrieval raises ROUGE-L from 0.422 to 0.438 and SBERT from 0.881 to 0.884, chiefly improving diagnostic consistency.

Qualitative Results: Fig. 2 shows an RDD test example. Retina-RAG leverages retrieved Grade-2 context to correctly identify microaneurysms, cotton-wool spots, and dot-and-blot hemorrhages consistent with Moderate NPDR; the macular assessment correctly reports no ME, with no hallucinated severe features.

Table 2. Component-wise ablation (ZS=Zero-shot, FT=Fine-tuned): (a) DR grading and ME detection; (b) report generation. Rows (Classifier+FT VLM, FLAIR) and (Retina-RAG) isolate the classifier-guided RAG module (retrieval OFF vs. ON).

(a) DR grading and ME detection

Config	Classifier	DR Severity (5-class)			ME (Binary)		
		Prec.	Rec.	F1	Prec.	Rec.	F1
ZS Qwen (Frozen VLM)	–	0.076	0.204	0.096	0.743	0.722	0.732
FT VLM	–	0.395	0.252	0.160	0.731	0.479	0.525
Classifier + ZS Qwen	FLAIR	0.667	0.462	0.470	0.910	0.911	0.910
	VGG16	0.537	0.396	0.411	0.941	0.941	0.938
	RBF-SVM	0.691	0.459	0.470	0.899	0.899	0.899
Classifier + FT VLM	FLAIR	**0.782**	0.752	0.698	0.881	0.876	0.878
	VGG16	0.444	0.568	0.487	0.850	0.861	0.846
	RBF-SVM	0.682	0.701	0.674	0.860	0.861	0.861
Retina-RAG (Ours)	FLAIR	0.778	**0.795**	**0.731**	**0.949**	**0.947**	**0.948**

(b) Report generation (NLP metrics)

Config	Classifier	BLEU-4 [32]	ROUGE-1 [33]	ROUGE-L [33]	SBERT [34]	Clin-BERT [35]
ZS Qwen (Frozen VLM)	–	0.014	0.175	0.163	0.764	0.921
FT VLM	–	0.114	0.342	0.333	0.840	0.960
Classifier + ZS Qwen	FLAIR	0.016	0.166	0.158	0.695	0.895
	VGG16	0.019	0.176	0.167	0.699	0.898
	RBF-SVM	0.016	0.170	0.162	0.700	0.896
Classifier + FT VLM	FLAIR	0.186	0.439	0.422	0.881	0.975
	VGG16	0.165	0.410	0.395	0.863	0.970
	RBF-SVM	0.186	0.419	0.412	0.871	0.971
Retina-RAG (Ours)	FLAIR	**0.194**	**0.461**	**0.438**	**0.884**	**0.975**

Cost impact: The dual-branch classifiers are fine-tuned via linear probing on a single GPU/TPU; Qwen2.5-VL-7B is LoRA-adapted (Unsloth) by training 1.23% of parameters (103M/8.4B) in 367 minutes at 10.5 GB peak memory. The full pipeline runs on a single RTX 1650 at $< \$0.0002$ per report [25] orders of magnitude below proprietary models ($\sim \$0.01$–$\$0.05$ per query) enabling scalable low-resource screening.

Ablation Studies: We ablate five configurations: (1) frozen VLM, (2) fine-tuned VLM, (3) classifier + frozen VLM, (4) classifier + fine-tuned VLM, and (5) full Retina-RAG with retrieval. The classifier drives the DR-grading gain, fine-tuning drives report quality, and retrieval adds the final diagnostic-consistency gain (isolated in Sec. 3.1). We also swap FLAIR for a VGG-16 [29] and a dual GoogleNet + ResNet-18 [28] configuration with an RBF-SVM [23] head, trained as the retinal classifier.

4 Conclusion

We present Retina-RAG, a modular framework for joint retinal diagnosis and long-form report generation that performs strongly while tuning only 1.23% of parameters. Its main limitation is that retrieval is conditioned on the predicted class, so classifier errors can propagate (quantified in Sec. 3.1); evaluation is also single-dataset, and NLP metrics measure proximity rather than clinical actionability. Future work targets cross-dataset generalization, expert factuality assessment, and prospective clinical validation.

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

References

1. Teo, Z.L. et al.: Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic review and meta-analysis. *Ophthalmology* **128**(11), 1580–1591 (2021)
2. Pan, J. et al.: Global, regional and national burden of blindness and vision loss attributable to diabetic retinopathy, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *Diabetes, Obesity and Metabolism* (2025). https://doi.org/10.1111/dom.16588
3. Tan, M., Le, Q.V.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 6105–6114 (2019)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (2016)
5. Humayun, M. et al.: Enhancing diabetic retinopathy classification using deep learning. *Digital Health* **9** (2023). https://doi.org/10.1177/20552076231203676
6. Abràmoff, M.D., Lavin, P.T., Birch, M., Shah, N., Folk, J.C.: Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. npj Digital Medicine **1**, 39 (2018). https://doi.org/10.1038/s41746-018-0040-6
7. Li, C. et al.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In: *NeurIPS Datasets and Benchmarks Track* (2023)
8. Gu, J. et al.: MedVH: Toward systematic evaluation of hallucination for large vision language models in the medical context. *Advanced Intelligent Systems* (2025). https://doi.org/10.1002/aisy.202500255
9. Chen, J. et al.: Detecting and evaluating medical hallucinations in large vision language models. In: *ICLR* (2025)
10. He, J. et al.: PeFoMed: Parameter efficient fine-tuning of multimodal large language models for medical imaging. *arXiv preprint arXiv:2401.02797* (2024)
11. Lewis, P. et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks.In: *Advances in Neural Information Processing Systems*, vol. 33,pp. 9459–9474 (2020)
12. Xia, P. et al.: RULE: Reliable multimodal RAG for factuality in medical vision language models. In: *EMNLP*, pp. 1081–1093 (2024)

13. Xia, P. et al.: MMed-RAG: Versatile multimodal RAG system for medical vision language models. In: *ICLR* (2025). arXiv:2410.13085
14. Abdalkader, M.: Retinal Disease Detection Dataset. Kaggle (2023). https://www.kaggle.com/datasets/mohamedabdalkader/retinal-disease-detection
15. Decencière, E. et al.: Feedback on a publicly distributed image database: the Messidor database. *Image Analysis and Stereology* **33**(3), 231–234 (2014). https://doi.org/10.5566/ias.1155
16. Porras, A.R. et al.: FLAIR: A foundation model for retinal image analysis. *arXiv preprint* arXiv:2501.09706 (2025)
17. Wilkinson, C.P., Ferris, F.L., Klein, R.E., et al.: Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology **110**(9), 1677–1682 (2003). https://doi.org/10.1016/S0161-6420(03)00475-5
18. American Academy of Ophthalmology: Diabetic Retinopathy Preferred Practice Pattern®. Ophthalmology **127**(1), P66–P145 (2020). https://doi.org/10.1016/j.ophtha.2019.09.025
19. Johnson, J. et al.: Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* **7**(3), 535–547 (2019)
20. Wang, P. et al.: Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint* arXiv:2409.12191 (2024)
21. Han, D., Han, M.: Unsloth: Efficient LLM fine-tuning (2024). https://github.com/unslothai/unsloth
22. Hu, E. et al.: LoRA: Low-rank adaptation of large language models. In: *ICLR* (2022)
23. Behera, M.K., Mishra, R., Ransingh, A., Chakravarty, S.: Prediction of different stages in diabetic retinopathy from retinal fundus images using radial basis function based SVM. Indian Journal of Science and Technology **13**(20), 2030–2040 (2020). https://doi.org/10.17485/IJST/v13i20.322
24. Dettmers, T. et al.: QLoRA: Efficient finetuning of quantized LLMs. In: *NeurIPS* (2023)
25. Hyperbolic Labs: Serverless inference pricing (2025). https://docs.hyperbolic.xyz/docs/hyperbolic-ai-inference-pricing
26. Bai, S. et al.: Qwen2.5-VL technical report. *arXiv preprint* arXiv:2502.13923 (2025)
27. Li, M. et al.: FFA-IR: Towards an explainable and reliable medical report generation benchmark. In: *Thirty-fifth Conference on NeurIPS Datasets and Benchmarks Track* (2021)
28. Butt, M.M., Iskandar, D.N.F.A., Abdelhamid, S.E., Latif, G., Alghazo, R.: Diabetic retinopathy detection from fundus images of the eye using hybrid deep learning features. Diagnostics **12**(7), 1731 (2022). https://doi.org/10.3390/diagnostics12071731
29. da Rocha, D.A., Ferreira, F.M.F., Peixoto, Z.M.A.: Diabetic retinopathy classification using VGG16 neural network. Research on Biomedical Engineering **38**, 761–772 (2022). https://doi.org/10.1007/s42600-022-00200-8
30. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF ICCV, pp. 10012–10022 (2021). https://doi.org/10.1109/ICCV48922.2021.00986
31. Ouyang, L., Wu, J., Jiang, X., et al.: Training language models to follow instructions with human feedback. In: NeurIPS, vol. 35, pp. 27730–27744 (2022)
32. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: A method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the ACL, pp. 311–318. ACL, Philadelphia (2002). https://doi.org/10.3115/1073083.1073135

33. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, pp. 74–81. ACL, Barcelona (2004)
34. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP), pp. 3982–3992. ACL, Hong Kong (2019). https://doi.org/10.18653/v1/D19-1410
35. Huang, K., Altosaar, J., Ranganath, R.: ClinicalBERT: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:1904.05342 (2019)
36. Huang, J.-H., Yang, C.-H.H., Liu, F., et al.: DeepOpht: Medical report generation for retinal images via deep models and visual explanation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 2442–2452 (2021)