MONASH UNIVERSITY

FACULTY OF INFORMATION TECHNOLOGY

From Prompts to Probes: Zero-Shot and Few-Shot Transfer of Foundation Models for Retinal Fundus Classification

FIT5128 –Final Thesis Paper

Student: Pugalenthi Magendran

Student ID: 31897800

Email: pmag0003@student.monash.edu

Supervisor: Dr Yasmeen George

Word Count (Main Thesis): 7900

TABLE OF CONTENT

Acknowledgments

I thank my supervisor, Dr Yasmeen George, for guidance and support throughout this project. I’m grateful to colleagues and collaborators for feedback, and to the providers of the datasets and open-source tools used in this work. I also appreciate the encouragement from friends and family during the thesis period. Any errors are my own.

Part I

Literature Review

Evaluating Foundation Model Adaptation for Retinal Disease Diagnosis: A Comparative Study of Zero-Shot Inference and Linear Probing with FLAIR

1 Introduction

Vision impairment is a major global health concern, affecting over 2.2 billion individuals worldwide, with at least 1 billion cases being preventable or unaddressed (World Health Organization, 2023). Among the primary contributors to visual disability are retinal diseases such as diabetic retinopathy (DR) and glaucoma, both imposing substantial global burdens. DR is a leading cause of vision loss among working-age adults, while glaucoma, a progressive optic neuropathy, is often asymptomatic until advanced stages. Timely and accurate diagnosis of these diseases is essential to prevent irreversible vision loss. However, growing demand for eye care services, combined with a projected 30% shortfall in ophthalmologist availability by 2035 in countries like the United States, has intensified interest in scalable, automated diagnostic solutions (Berkowitz et al., 2023).

Artificial intelligence (AI), particularly convolutional neural networks (CNNs), shows strong potential in retinal diagnostics. Landmark studies have demonstrated expert-level performance in detecting diabetic retinopathy (Gulshan et al., 2016; Ting et al., 2017). However, these models require large-annotated datasets and are often disease-specific, limiting generalizability. This reliance on labelled data limits deployment in low-resource settings. Moreover, one of the key challenges in clinical AI deployment is ensuring that models trained in one setting generalize effectively to other populations, devices, and disease presentations. Without this robustness, even highly accurate models may underperform in new environments, a critical concern in ophthalmology, where fundus image characteristics vary with acquisition protocols and demographics (Rashidisabet et al., 2023).

To address these limitations, the AI research community has shifted toward foundation models, large-scale, pre-trained architectures capable of capturing generalizable representations across multiple tasks. In medical imaging, several domain-specific foundation models have emerged, including RETFound, RetiZero, ViLReF, and FLAIR (Zhou et al., 2023; Wang et al., 2024; Yang et al., 2024; Silva-Rodriguez et al., 2025). These models leverage transfer learning on diverse retinal datasets for tasks like classification, segmentation, and progression prediction. FLAIR (Foundation Language-Image Model of the Retina) is a vision-language foundation model

developed for retinal diagnostics (Silva-Rodriguez et al., 2025). It was pre-trained on over 288,000 fundus images from 38 open-access datasets, paired with expert-written disease descriptions. Using contrastive learning, FLAIR maps images and text into a shared embedding space, enabling zero-shot inference, without retraining. Leveraging textual prompts alone, it holds strong potential for deployment in low-resource settings with limited annotations.

The aim of this literature review is to critically evaluate the use of foundation models in retinal disease diagnosis, with a particular focus on the FLAIR vision-language model. It investigates two prominent adaptation strategies, zero-shot inference and linear probing, and examines their performance on unseen benchmark datasets. The review identifies current limitations in data efficiency and generalization, highlights gaps in model evaluation, and synthesizes insights to guide the deployment of foundation models in underrepresented ophthalmic conditions. While FLAIR demonstrates strong performance on several benchmark tasks, a key challenge remains: how should clinicians and researchers adapt this model to new diagnostic settings with limited labelled data? One approach is to use FLAIR's zero-shot inference capabilities directly. Another is to employ lightweight adaptation methods, such as linear probing (LP), which trains a simple classifier on frozen image embeddings. Each strategy presents trade-offs in terms of data efficiency, diagnostic accuracy, and computational cost.

Although prior work has reported FLAIR's zero-shot performance on held-out datasets such as REFUGE and Messidor, no comprehensive evaluation exists comparing zero-shot inference with linear probing on these benchmarks (Silva-Rodriguez et al., 2025). REFUGE and Messidor are particularly well-suited for assessing model generalization, as they were explicitly excluded during FLAIR's pretraining. These datasets are widely used in ophthalmic AI; Messidor serves as a benchmark for diabetic retinopathy detection and REFUGE as a framework for glaucoma assessment and optic disc/cup segmentation (Decencière et al., 2014; Orlando et al., 2020). While DR and glaucoma are common, this study simulates low-resource settings using constrained training data to reflect conditions in rare or underrepresented retinal diseases and inform low-data scenarios.

Section 2 reviews existing work on AI in ophthalmology, the emergence of vision-language foundation models, adaptation techniques such as zero-shot inference and linear probing, challenges in generalization, interpretability, and deployment. Section 3 synthesizes the state of the art and clearly defines the research gap and questions. Section 4 outlines the proposed research plan, including data collection, preprocessing, model development, evaluation metrics, risk, and ethical considerations. Finally, Section 5 concludes the report by summarizing the contribution and its relevance to the broader field of medical AI, particularly in guiding foundation model deployment in low-resource, real-world clinical environments.

2 Substantive Literature Review

2.1 Advancements and Limitations of AI in Ophthalmology

2.1.1 Early Development of CNN-based Tools

Convolutional neural networks (CNNs) marked a breakthrough in ophthalmic AI by achieving expert-level performance in detecting diabetic retinopathy (DR) from fundus photographs, with early models demonstrating sensitivity and specificity comparable to ophthalmologists (Gulshan et al., 2016). This led to regulatory milestones such as IDx-DR, the first FDA-approved autonomous AI system for DR screening (Abràmoff et al., 2018), signalling the transition from research to clinical application. Subsequent studies extended CNNs to other retinal diseases, including glaucoma and age-related macular degeneration, using large, diverse datasets across regions, underscoring their scalability for population-level screening (Ting et al., 2017). However, supervised CNNs face major limitations: they require extensive expert-labelled data, are often task-specific, and struggle with domain shift, showing reduced generalizability across different imaging settings and populations.

While open-access datasets have mitigated data scarcity in ophthalmic AI, generalization remains a key challenge. Models trained on datasets like EyePACS often underperform on external benchmarks such as Messidor or IDRiD due to variations in imaging conditions, demographics, and disease patterns (Gulshan et al., 2016; Nadeem et al., 2022). Although IDRiD offers valuable annotations and broader representation, its geographic specificity highlights the limitations of localized training data (Porwal et al., 2018). Moreover, most supervised CNNs are tailored to

narrow tasks, failing to capture the complexity of comorbid or heterogeneous retinal diseases. These constraints have accelerated interest in more adaptable approaches, including transfer learning, self-supervised learning, and foundation models, discussed in the following sections.

2.1.2 Generalization and Deployment Challenges in Ophthalmic AI

Despite early advances, real-world deployment of AI in ophthalmology remains hindered by challenges in generalization, dataset diversity, and clinical integration. Models that perform well in internal validation often degrade in new settings due to domain shift, differences in imaging devices, protocols, disease prevalence, and patient demographics (Nadeem et al., 2022). For example, the IDRiD dataset offers detailed annotations but is limited by its geographic and technical uniformity, reducing its applicability across diverse populations (Porwal et al., 2018). These constraints underscore the need for more representative datasets and robust learning strategies, driving growing interest in scalable approaches such as transfer learning, self-supervised learning, and foundation models.

Benchmark datasets in medical imaging often contain hidden stratification and label noise, resulting in overstated validation accuracy but limited clinical reliability. These challenges, well-documented in radiology, are equally relevant in ophthalmology, where subtle features and inter-observer variability hinder consistent annotation (Oakden-Rayner et al., 2020). Public ophthalmic datasets are typically unimodal, small, and class-imbalanced, restricting the development of models suited for real-world multimodal inputs like fundus images, OCT scans, and clinical data (Wang et al., 2024). Additionally, the absence of interpretability and uncertainty quantification in current models undermines clinical trust. Recent assessments of multimodal systems integrating imaging with large language models highlight deficiencies in transparency and consistency, underscoring the need for more explainable and robust AI frameworks (Peng et al., 2023). Overcoming these limitations calls for a paradigm shift in how AI models are trained, validated, and deployed in ophthalmology.

2.2 Foundation Models in Ophthalmology

2.2.1 Paradigm Shift: From Task-Specific Models to Foundational Architectures

Foundation models represent a major advancement in AI, offering strong generalization by learning from large-scale unlabelled data through self- or weakly supervised methods (Bommasani et al., 2021). Unlike task-specific CNNs, they capture semantic and structural features that support adaptation to diverse diagnostic tasks with minimal labelled data, an asset in ophthalmology, where annotated datasets are limited by privacy and cost (Shi et al., 2024). These models show improved robustness to domain shift and strong performance in low-data scenarios (Azizi et al., 2021; Nguyen et al., 2023). Architectures such as Vision Transformers and multimodal encoders, trained with contrastive learning, align visual and textual inputs (Qiu et al., 2023; Kumar & Marttinen, 2024). Their adaptability via linear probing or few-shot learning makes them suitable for clinical settings with heterogeneous imaging and sparse labels (Pachetti & Colantonio, 2024; Li et al., 2022). Additionally, uncertainty estimation and interpretability techniques help build clinical trust, as highlighted in evaluations like the HRDC Challenge (Qian et al., 2024; Zou et al., 2023).

2.2.2 Notable Ophthalmic Foundation Models

Several foundation models have recently been developed specifically for ophthalmology, each addressing key limitations of earlier AI systems:

- **RETFound** was the first publicly released foundation model trained on 1.6 million unlabelled fundus images using self-supervised learning. It demonstrated strong performance in diabetic retinopathy (DR) and age-related macular degeneration (AMD) classification tasks, especially in low-label environments, highlighting the importance of scale and pretraining in ophthalmic diagnostics (Zhou et al., 2023).

- **UrFound** expanded on RETFound by incorporating a Vision Transformer (ViT) backbone and contrastive objectives. Trained across both fundus and OCT modalities, its task-agnostic design was validated on benchmarks such as REFUGE and IDRiD, supporting various clinical use cases with minimal fine-tuning (Yu et al., 2024).

- **RetiZero** introduced zero-shot inference for ophthalmology using a vision-language model trained with contrastive learning. It enables natural language prompting to infer disease characteristics without requiring labelled training data, offering a practical solution for rare or emerging conditions (Wang et al., 2024).

- **RET-CLIP** adapted CLIP's vision-language framework with domain-specific medical supervision, aligning retinal images with curated disease descriptions to improve zero-shot inference accuracy in ophthalmic contexts (Du et al., 2024).

- **ViLReF** focused on aligning fundus images with ophthalmologist-written clinical reports. By grounding model outputs in free-text expert descriptions, it enhances interpretability and supports explainable AI in retinal diagnostics (Yang et al., 2024).

- **EyeFound** and **VisionUnite** explored multimodal learning by integrating OCT, fundus images, and structured clinical metadata across institutional datasets. These models prioritize privacy-preserving training and emphasize cross-site generalizability in real-world settings (Shi et al., 2024; Li et al., 2024).

- **EyeCLIP** is a large-scale vision-language foundation model developed to support multimodal alignment across various retinal imaging modalities. It enables zero-shot inference, few-shot rare disease diagnosis, and visual question answering, advancing generalist capabilities in retinal AI (Shi et al., 2024).

- **FLAIR** represents a specialized vision-language foundation model for retinal imaging. Using contrastive learning with expert-derived clinical prompts, it facilitates zero-shot inference and interpretable report generation, exemplifying the shift toward flexible and clinically grounded AI tools (Silva-Rodríguez et al., 2025).

These models collectively mark a transition from narrow, task-specific architectures to scalable, adaptable, and clinically trustworthy AI systems in ophthalmology.

2.2.3 Challenges and Opportunities Ahead

Despite their potential, foundation models in ophthalmology face key challenges. A major issue is domain mismatch, general-purpose models like CLIP or ChatGPT often underperform in retinal imaging due to the absence of ophthalmology-specific priors, failing to capture fine-grained pathological features in fundus or OCT images (Sevgi et al., 2025). Additionally, the lack of large, well-curated ophthalmic datasets, driven by privacy concerns, limited collaboration, and high annotation costs, hampers model generalizability and increases bias risk. While solutions like synthetic augmentation and multi-institutional training exist, they remain underutilized in current development efforts (Shi et al., 2024; Qiu et al., 2023).

Interpretability and trust remain critical barriers to clinical adoption of foundation models, as many lack mechanisms for uncertainty estimation and transparent decision-making, essential for high-stakes medical use. Incorporating explainability, fairness audits, and calibration is vital for safe, equitable deployment (Peng et al., 2023; Shi et al., 2024; Zou et al., 2023). Future progress hinges on advancing multimodal architectures that combine imaging, clinical text, and health records; refining lightweight adaptation methods like linear probing and few-shot learning; and establishing robust benchmarks across diverse settings. Addressing these gaps is key to making foundation models clinically reliable, scalable, and equitable in retinal diagnostics.

2.3 FLAIR: A Vision-Language Model for Retinal Diagnosis

2.3.1 Pretraining and Architectural Design

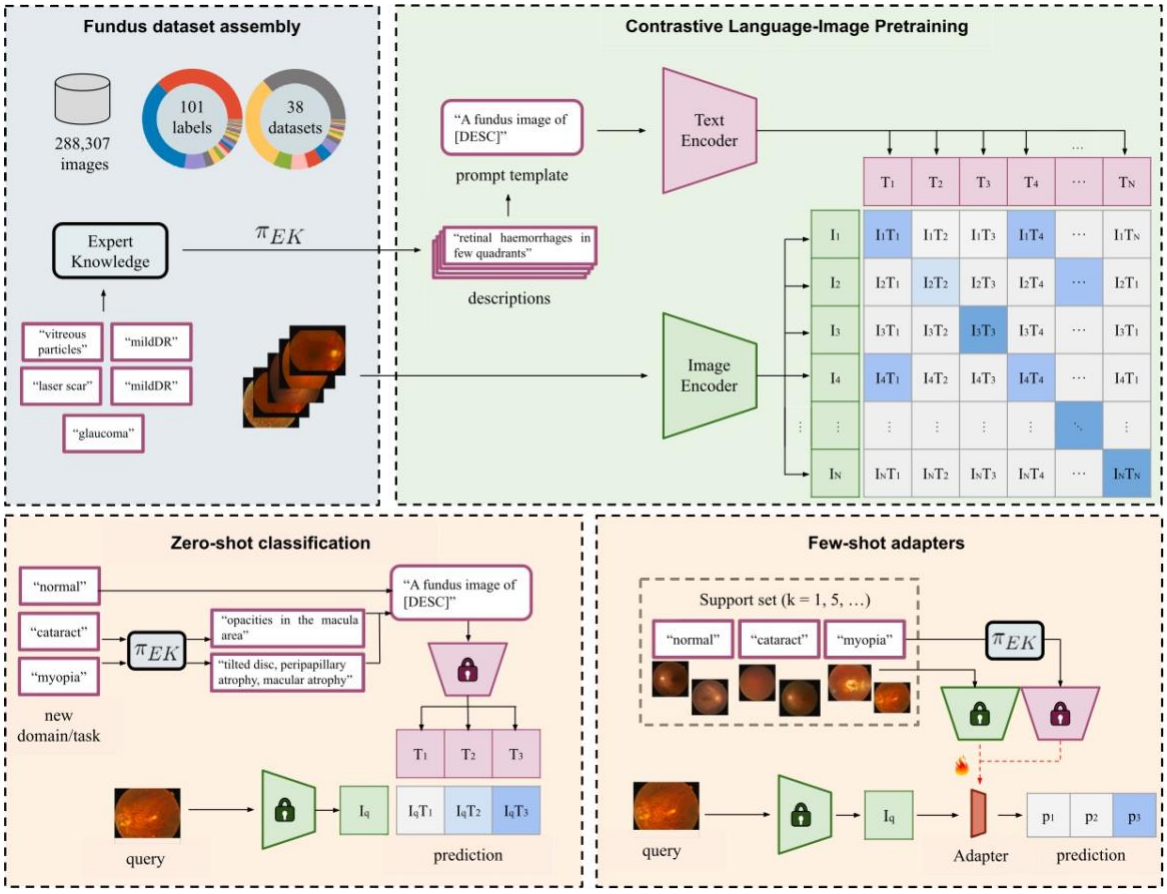


Figure 1

FLAIR's architecture and training framework. The model uses a dual-encoder (image and text) design trained with contrastive learning on over 288,000 fundus images and expert-aligned text prompts. It supports zero-shot and few-shot classification using prompt-based inference or lightweight adaptation. Adapted from Silva-Rodríguez, J., Liu, H., Ayhan, M. S., Wang, J., Zhang, Y., Yang, D., ... & Müller, H. (2025). FLAIR: Foundation Language-Image Model of the Retina (arXiv:2308.07898). https://arxiv.org/abs/2308.07898

Building on the broader foundation model paradigm (Bommasani et al., 2021), FLAIR is a vision-language foundation model specifically developed for retinal fundus image analysis (Silva-Rodríguez et al., 2025). It employs a dual-encoder architecture comprising a ResNet-50 visual backbone and a BioClinicalBERT language model, aligned through contrastive learning. In contrast to earlier ophthalmic models such as RETFound and UrFound, which relied solely on image-based supervision, FLAIR incorporates structured domain knowledge through expert-crafted textual prompts. This integration enables the model to associate clinical language with visual pathology effectively.

FLAIR was pretrained on over 288,000 images sourced from 38 diverse fundus datasets, using categorical labels mapped to domain-specific textual descriptions (Silva-Rodríguez et al., 2025). This design facilitates scalable supervision across numerous disease categories and enhances the model's ability to generalize to new tasks and domain shifts. Experimental results indicate that FLAIR outperforms both task-specific convolutional neural networks and general-purpose models such as CLIP, particularly in few-shot and zero-shot settings (Silva-Rodríguez et al., 2025).

2.3.2 Contrastive Learning and Zero-Shot Inference

FLAIR's most notable contribution lies in its ability to perform zero-shot inference. During inference, it compares a retinal image's embedding to a set of domain-knowledge textual prompts, for instance, "only a few microaneurysms are present", using cosine similarity (Silva-Rodríguez et al., 2025). This mechanism allows the model to infer disease presence without task-specific retraining, offering clinical flexibility, especially in the context of rare or previously unseen conditions.

In benchmarking experiments, FLAIR consistently outperformed generalist vision-language models such as CLIP and BiomedCLIP and demonstrated strong generalization under domain shifts (Silva-Rodríguez et al., 2025). While supervised convolutional neural networks (CNNs) retained an advantage in high-data or fine-grained diagnostic tasks, FLAIR achieved competitive or superior performance in low-data and few-shot scenarios. Lightweight adaptation using linear probing further enhanced its performance, underscoring its practical utility for real-world clinical deployment (Silva-Rodríguez et al., 2025).

2.3.3 Dataset Holdout Strategy

To rigorously evaluate its generalization capacity, FLAIR's developers adopted a dataset holdout strategy in which public datasets such as Messidor (for diabetic retinopathy; Decencière et al., 2014) and REFUGE (for glaucoma and optic disc segmentation; Orlando et al., 2020) were excluded from the pretraining phase and used exclusively for downstream testing (Silva-Rodríguez et al., 2025). This design ensures that performance metrics reflect true generalization rather than memorization of training data.

The approach aligns with recent evaluation recommendations in medical AI, which emphasize the importance of out-of-distribution testing to ensure robust and fair assessment of foundation models (Jin et al., 2024). FLAIR's strong performance on held-out datasets supports its readiness for deployment in clinical environments where prior exposure to specific data distributions cannot be assumed (Silva-Rodríguez et al., 2025).

2.4 Adaptation Strategies: Zero-Shot vs. Linear Probing

2.4.1 Zero-Shot Learning and Lightweight Adaptation

Adaptation strategies are key to leveraging foundation models like FLAIR for downstream tasks. Zero-shot inference enables predictions without retraining by aligning image features with textual prompts, using metrics like cosine similarity between fundus images and expert-crafted descriptions (Radford et al., 2021; Silva-Rodríguez et al., 2025). This approach allows rapid adaptation without labelled data, making it practical for diverse clinical scenarios. However, studies such as ViLReF highlight that while zero-shot performance is promising, it often underperforms compared to fully supervised models, especially for rare or subtle diagnostic categories (Yang et al., 2024).

Linear probing trains a lightweight classifier, such as logistic regression, on frozen pretrained features using a small labelled dataset. This approach leverages the general representations of foundation models while adding minimal supervision. In FLAIR, linear probing yields significant performance gains over zero-shot inference, improving AUC by 5–12% depending on the task (Silva-Rodríguez et al., 2025). Benchmarks like the HRDC Challenge highlight its effectiveness, showing that linear probing offers a strong balance between accuracy and efficiency, often rivaling full fine-tuning with lower computational cost (Qian et al., 2024).

Adapter-based learning offers scalable adaptation without altering the full model backbone. Techniques like CLIP-Adapter enhance few-shot classification by inserting residual adapters into the visual or language branches, blending frozen and task-specific knowledge efficiently (Gao et

al., 2024). In ophthalmology, FundusAdapter uses hierarchical modules with gated cross-attention and memory to capture global and local features, achieving strong few-shot performance on complex retinal lesions (Chang et al., 2025). These strategies provide a practical middle ground between zero-shot and full fine-tuning, enabling cost-effective deployment of foundation models in clinical settings.

2.4.2 Performance Trade-offs Between Zero-Shot and Linear Probing

Zero-shot approaches excel in generalization, particularly under domain shift. For instance, FLAIR has demonstrated strong performance on external datasets without retraining, showing robustness in unseen populations. However, this flexibility often comes at the expense of fine-grained accuracy, zero-shot inference struggles to distinguish subtle severity levels of diseases such as diabetic retinopathy or glaucoma compared to supervised models (Yu et al., 2023). In contrast, linear probing leverages modest amounts of labelled data to refine decision boundaries. RETFound, for example, saw a 12.7% improvement in AUC when linear probing was applied on external test sets, and FLAIR similarly demonstrated consistent gains in evaluation benchmarks like HRDC (Qian et al., 2024; Zhou et al., 2023). Ultimately, the choice between zero-shot and linear probing hinges on the deployment context. Zero-shot learning offers scalable, low-labor solutions when annotations are scarce, while linear probing provides a cost-effective path to improved accuracy in clinically sensitive scenarios. Emerging approaches such as adapter tuning and few-shot calibration may offer hybrid solutions that bridge this gap, balancing performance with adaptability in real-world medical settings.

2.5 Benchmark Studies and Evaluation Gaps

2.5.1 MM-Retinal and HRDC Findings

Benchmark studies have been instrumental in assessing foundation models for fundus image analysis. The MM-Retinal benchmark and HRDC Challenge are among the most prominent efforts. MM-Retinal introduced a multimodal evaluation suite spanning CFP, FFA, and OCT, with its foundation model KeepFIT achieving strong zero- and few-shot performance via image-guided prompt revision (Wu et al., 2024a). MM-Retinal V2 further improved cross-task transferability

through hybrid prompting and expert supervision (Wu et al., 2025). However, reliance on proprietary datasets limits reproducibility and broader comparison. In contrast, RETFound emphasizes transparency by releasing pretrained weights, code, and benchmarking results on public datasets such as MESSIDOR-2, IDRiD, and APTOS-2019 (Zhou et al., 2023), setting a reproducibility standard in ophthalmic AI (Silva-Rodríguez et al., 2025).

The HRDC Challenge, presented at CGI 2023, evaluated hypertensive retinopathy classification using 2,000 fundus images, with submissions ranging from conventional models to foundation model adaptations like FLAIR (Qian et al., 2024; Silva-Rodríguez et al., 2025). Linear probing over pretrained embeddings achieved strong results, with a top Kappa score of 0.4154 and specificity of 0.8444 (Qian et al., 2024). However, the dataset's limited diagnostic scope, single-site origin, and uniform imaging conditions restrict its clinical generalizability. Collectively, MM-Retinal, RETFound, and HRDC illustrate progress in ophthalmic AI benchmarking but also highlight the need for standardized, diverse, and publicly accessible evaluation frameworks.

2.5.2 Controlled Evaluation on REFUGE and Messidor: Progress and Remaining Gaps

Foundation models like FLAIR have advanced generalization in fundus image analysis, notably through a domain shift evaluation protocol that excluded public benchmarks, REFUGE, Messidor, and FIVES, from pretraining, using them solely for downstream testing to fairly assess zero-shot and linear probing performance on out-of-distribution data (Silva-Rodríguez et al., 2025; Decencière et al., 2014; Orlando et al., 2020). However, evaluations across FLAIR and MM-Retinal remain inconsistent due to variations in prompt formulation, label granularity, and dataset partitioning. While FLAIR provides dataset-specific metrics, the lack of standardized benchmarking pipelines limits reproducibility and complicates meaningful cross-model comparisons (Silva-Rodríguez et al., 2025).

Adaptation task framing varies widely across studies, complicating direct comparisons. For example, MM-Retinal V2 introduces multi-label, multi-modal tasks and hybrid prompting (Wu et al., 2025), making it difficult to benchmark against conventional methods like zero-shot inference or linear probing. Additionally, the combined use of techniques such as adapter tuning, prompt engineering, and expert-guided supervision, especially on canonical datasets like Messidor and

REFUGE, makes it hard to isolate their individual contributions (Decencière et al., 2014; Orlando et al., 2020). To address these issues, our study introduces a reproducible evaluation framework using standardized public datasets, fixed partitions, consistent prompts, and class-balanced protocols. This setup enables transparent comparisons between adaptation strategies and supports the development of fair, clinically meaningful benchmarks for ophthalmic foundation models.

2.6 Simulating Rare Disease Settings via Low-Data Evaluation

2.6.1 Motivation for Using DR/Glaucoma with Limited Data

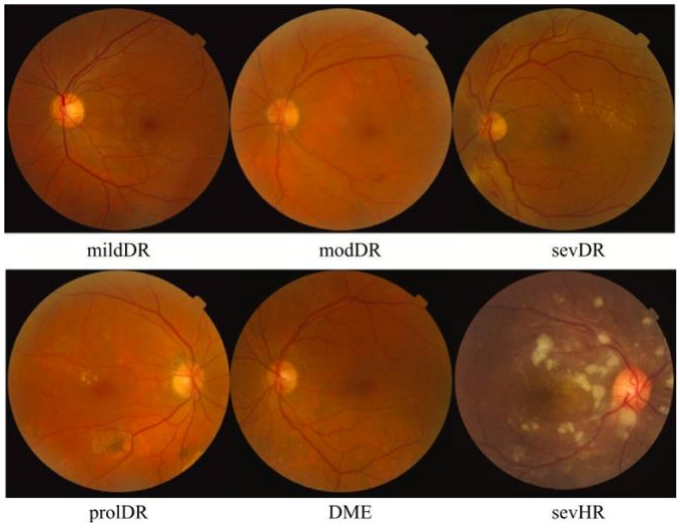


Figure 2

Representative fundus images across diabetic retinopathy (DR) stages and related retinal conditions. These illustrate the progressive visual complexity that challenges annotation and classification. Adapted from Silva-Rodríguez, J., Liu, H., Ayhan, M. S., Wang, J., Zhang, Y., Yang, D., ... & Müller, H. (2025). FLAIR: Foundation Language-Image Model of the Retina (arXiv:2308.07898). https://arxiv.org/abs/2308.07898

Diabetic retinopathy (DR) and glaucoma are leading causes of global vision loss, where early detection is critical to prevent irreversible damage (Ting et al., 2017; Bourne et al., 2021). Although prevalent, their early and atypical presentations, along with underrepresentation in certain populations, pose challenges for obtaining diverse, well-annotated training data (Holmberg et al., 2020). Public datasets like Messidor (DR) and REFUGE (glaucoma) often lack variability in imaging devices, ethnicity, and disease subtypes, limiting model generalizability. To address this, foundation models such as FLAIR deliberately excluded these datasets during pretraining,

using them solely for downstream evaluation to simulate low-data and domain shift scenarios (Silva-Rodríguez et al., 2025).

Simulating low-resource conditions with DR and glaucoma offers a clinically relevant and reproducible way to assess foundation model robustness. Studies show that self-supervised and few-shot methods can achieve high performance using limited data, for instance, accurate DR classification with just 5–25% of labelled data (Holmberg et al., 2020), or as few as 20 samples per class (Pan et al., 2021; Murugappan et al., 2022). Additionally, meta-learning frameworks have used DR and glaucoma to improve generalization to rare retinal diseases (Gao et al., 2023). Their diagnostic complexity and availability in public datasets make them ideal proxies for evaluating data-efficient adaptation in ophthalmic AI.

2.6.2 Relevance to Real-World Deployment Scenarios

Many clinical settings, particularly in low-resource environments, lack the labelled data, computational resources, and specialist access required for fully supervised AI systems (Ting et al., 2019). In such cases, data-efficient strategies like zero-shot learning and linear probing are essential for practical deployment.

Recent studies show that clinically relevant features can be learned with limited supervision. Self-supervised retinal thickness prediction improved DR classification using unlabelled images (Holmberg et al., 2020), while meta-learning enabled rapid adaptation to various retinal tasks (Wang et al., 2024). In-context learning with models like Gemini 1.5 Pro achieved competitive performance without retraining or large labelled datasets (Ayhan et al., 2025).

A few-shot ensemble method also proved effective in classifying rare fundus diseases through hierarchical feature generalization (Gao et al., 2023). These findings emphasize the value of simulating low-data scenarios to ensure the reliability and scalability of foundation models in underserved settings.

3 Summary of the State of the Art

Foundation models are transforming ophthalmic AI by addressing the limitations of traditional CNNs, which depend on large, annotated datasets and suffer from poor generalizability due to domain shift (Gulshan et al., 2016; Ting et al., 2017). Recent models like RETFound, RetiZero, ViLReF, and FLAIR leverage large-scale unlabelled data and contrastive learning to enable zero- and few-shot classification across diverse retinal tasks (Zhou et al., 2023; Wang et al., 2024; Yang et al., 2024; Silva-Rodríguez et al., 2025). FLAIR, for instance, employs a CLIP-inspired dual-encoder that aligns fundus images with expert prompts and excludes benchmarks like Messidor and REFUGE during pretraining to support robust domain-shift evaluation (Radford et al., 2021).

However, models like MM-Retinal V2 rely on internal datasets, limiting reproducibility (Wu et al., 2025), while synthetic augmentation approaches like SynFundus face questions about clinical realism. Emerging research in multimodal self-supervised learning and distributed training highlights the urgent need for reproducible, scalable, and equitable evaluation frameworks in ophthalmic AI (Sükei et al., 2024; Gholami et al., 2025).

3.1 Identified Research Gaps

Gap 1: Absence of Controlled Comparison Between Zero-Shot and Linear Probing for FLAIR

While FLAIR shows strong performance in both zero-shot and linear probing settings, no study has directly compared these strategies using standardized public datasets like Messidor and REFUGE (Decencière et al., 2014; Orlando et al., 2020; Silva-Rodríguez et al., 2025). Variations in prompt design, dataset splits, and task framing across prior work hinder reproducibility and make direct comparisons difficult. This gap limits our understanding of which adaptation strategy performs best under clinically realistic constraints.

Gap 2: Lack of Reproducible Low-Data Benchmarks for Ophthalmology

Although DR and glaucoma are globally prevalent, early-stage, atypical, or comorbid cases are underrepresented in clinical datasets, creating low-data conditions that hinder model robustness. Simulated down sampling enables reproducible evaluation, and methods like few-shot learning and contrastive pretraining show promise (Holmberg et al., 2020; Pan et al., 2021; Gao et al., 2023). However, foundation models like FLAIR remain untested under these constraints on standardized benchmarks.

Gap 3: Deployment Trade-offs Between Zero-Shot and Adapted Models Remain Unclear

While foundation models offer scalability, their clinical reliability under zero-shot inference remains uncertain, particularly in low-resource settings. Prior work, including RETFound and MM-Retinal V2, highlights the benefits of linear probing, yet vision-language models like FLAIR remain underexplored in this regard (Zhou et al., 2023; Wu et al., 2025). This gap is critical where data, compute, and expert oversight are limited (Ting et al., 2019). To address this, we present the first controlled benchmark comparing zero-shot and linear probing for FLAIR on standardized datasets, Messidor and REFUGE, for DR and glaucoma classification (Decencière et al., 2014; Orlando et al., 2020). Simulated low-data settings via down sampling reveal key adaptation trade-offs, guiding practical deployment in ophthalmology.

3.2 Research Questions

RQ1: How effectively can the FLAIR vision-language model classify diabetic retinopathy and glaucoma in a zero-shot setting?

RQ2: Can linear probing improve FLAIR's classification performance for diabetic retinopathy and glaucoma compared to zero-shot inference?

RQ3: What are the performance trade-offs between zero-shot inference and linear probing when adapting the FLAIR model to retinal disease diagnosis tasks?

4 Research Project Plan

4.1 Project Objective

The overall aim of this project is to evaluate and compare the performance of zero-shot inference and linear probing as adaptation strategies for deploying the FLAIR foundation model in retinal disease diagnosis, particularly in low-data clinical settings.

The specific objectives of this research are:

1. To train and evaluate a foundation model to classify diabetic retinopathy.
2. To train and evaluate a foundation model to classify glaucoma.
3. To implement a comprehensive study and analysis comparing performance across the two diseases, particularly under low-data constraints.

4.2 Methodology Overview

This study adopts a structured experimental pipeline consisting of four major components, each detailed in Sections 4.2.1 to 4.2.4:

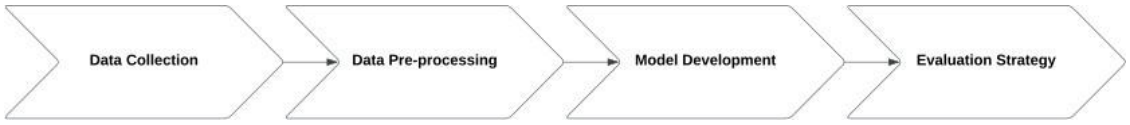


Figure 3

1. **Data Collection**:
   Involves sourcing and preparing annotated fundus images from public datasets like Messidor and REFUGE, and generating constrained training subsets at 1%, 5%, 10%, and 20% to simulate low-data clinical environments.
2. **Pre-processing**:
   Covers image resizing, normalization, label harmonization, prompt construction for zero-shot inference, and data augmentation strategies tailored for linear probing.

3. **Model Development**:
   Configures the FLAIR model for two adaptation strategies: zero-shot inference (using prompt-based similarity) and linear probing (training a classifier on frozen embeddings). Includes comparative benchmarking against models like ViT-ImageNet, RETFound, and FundusAdapter.
4. **Evaluation**:
   Encompasses training procedures, performance metric computation (e.g., AUROC, sensitivity), statistical significance testing, and reproducibility tracking through random seed control and code release.

4.2.1 Data Collection

This project uses publicly available, de-identified datasets that require no additional ethics approval:

- Messidor (Decencière et al., 2014): Contains 1,200 color fundus photographs labeled with diabetic retinopathy severity scores, ranging from no DR to proliferative DR.

- REFUGE (Orlando et al., 2020): A benchmark dataset of 1,200+ fundus images annotated for glaucoma diagnosis, including expert-verified cup-to-disc ratios.

These datasets were explicitly excluded from FLAIR's pretraining (Silva-Rodríguez et al., 2025), making them ideal for out-of-distribution evaluation. Training subsets will be sampled at 1%, 5%, 10%, and 20% to simulate low-resource settings.

The IDRiD dataset may be optionally included for supplementary lesion-level analysis, though its geographic and device homogeneity limits its suitability as a standalone generalization benchmark (Porwal et al., 2018). Messidor and REFUGE remain the primary datasets for adaptation performance comparison.

4.2.2 Data Pre-processing

Preprocessing was implemented to ensure compatibility with FLAIR's architecture and to enhance model generalization under low-data conditions.

All fundus images were resized to 224×224 pixels and normalized using the ImageNet mean and standard deviation, consistent with prior deep learning practices in ophthalmic image classification (Islam et al., 2022; Silva-Rodríguez et al., 2025). For zero-shot inference, clinical prompts were constructed in natural language, for instance, "presence of referable diabetic retinopathy", following contrastive prompt engineering protocols (Radford et al., 2021). Labels for diabetic retinopathy were binarized, and glaucoma annotations adhered to REFUGE expert diagnoses. To address class imbalance in low-data regimes, oversampling and under sampling strategies were employed (Das & Walia, 2024). Data augmentation, including random flipping, rotation, and brightness adjustments, was applied exclusively in the linear probing setup to test its effect on generalization. All preprocessing was conducted using Python libraries such as OpenCV, Pillow, and torch vision, with Weights & Biases used for experiment tracking to ensure reproducibility (Holmberg et al., 2020).

4.2.3 Model Development

This study evaluates two adaptation strategies using the pretrained FLAIR model to classify diabetic retinopathy and glaucoma.

Strategy 1: Zero-Shot Inference

FLAIR embeds a fundus image using its vision encoder (Vision Transformer, ViT) and compares it against disease-related text prompt embeddings generated by its language encoder (BioGPT). Classification is determined based on cosine similarity between the image and text embeddings. To enhance robustness, multiple semantically aligned prompts, for instance, "presence of referable diabetic retinopathy" will be used for each class. This inference method does not require additional training data, making it particularly suitable for deployment in data-scarce clinical environments (Silva-Rodríguez et al., 2025).

Strategy 2: Linear Probing

In this approach, FLAIR's vision encoder is frozen to preserve learned representations, and a lightweight L2-regularized logistic regression classifier is trained using the extracted image embeddings on a small labelled dataset. The language encoder is not utilized. This method allows task-specific adaptation with minimal supervision. Model training will use stochastic gradient descent with early stopping, and evaluation will follow a 5-fold stratified cross-validation strategy to ensure robustness (Silva-Rodríguez et al., 2025).

FLAIR's dual-encoder architecture is trained via contrastive learning to align ophthalmic image features with corresponding clinical language descriptions. This architecture enables zero-shot transfer and flexible downstream adaptation, positioning it as a promising foundation model for real-world ophthalmic AI (Silva-Rodríguez et al., 2025).

To benchmark FLAIR, this study compares three models: ViT-ImageNet as a task-agnostic baseline, RETFound for its self-supervised adaptation to fundus imaging (Zhou et al., 2023), and FundusAdapter for its modular design supporting low-data transferability (Chang et al., 2025). All models will be evaluated using consistent splits, preprocessing, and label definitions to ensure fair and reproducible comparison.

Computational Infrastructure

All experiments will be conducted on Monash University's M3 Massive HPC cluster, which offers GPU acceleration, high-throughput scheduling, and secure storage. Its scalability supports efficient training, cross-validation, and management of datasets, models, and logs for vision-language tasks like FLAIR (Goscinski et al., 2014).

4.2.4 Evaluation Strategy

To rigorously evaluate zero-shot inference and linear probing, this study adopts a comprehensive framework combining quantitative, qualitative, and statistical analyses. Core performance metrics include AUROC, accuracy, sensitivity, specificity, precision, and F1-score (Sokolova & Lapalme, 2009). Embedding separability will be visualized using t-SNE or PCA (van der Maaten & Hinton, 2008), and interpretability explored via Grad-CAM or similar tools (Selvaraju et al., 2017). Statistical significance will be assessed using paired t-tests or Wilcoxon signed-rank tests, with 95% confidence intervals estimated through bootstrapping or analytical methods (Efron & Tibshirani, 1993). Effect sizes will be reported using Cohen's d (Cohen, 2013). All experiments will be repeated with three random seeds to account for variability, and results visualized using seaborn and matplotlib (Waskom, 2021). Code and configurations will be made publicly available via GitHub under an open-source license to ensure reproducibility.

4.3 Expected Outcomes

This study presents the first controlled evaluation of zero-shot versus linear probing strategies for the FLAIR foundation model on standardized datasets, Messidor for diabetic retinopathy (DR) and REFUGE for glaucoma (Decencière et al., 2014; Orlando et al., 2020). By simulating low-data clinical conditions, it assesses whether lightweight adaptation improves performance over zero-shot inference in ophthalmic vision-language models. The findings will clarify adaptation trade-offs and generalizability for fundus image classification, addressing two leading causes of global vision loss (Ting et al., 2017; Bourne et al., 2021). Additionally, the project introduces a reproducible benchmark, complete with dataset splits, evaluation code, and prompt templates, to support future ophthalmic AI research and enable deployment in resource-limited settings (Ting et al., 2019; Silva-Rodríguez et al., 2025).

4.4 Timeline

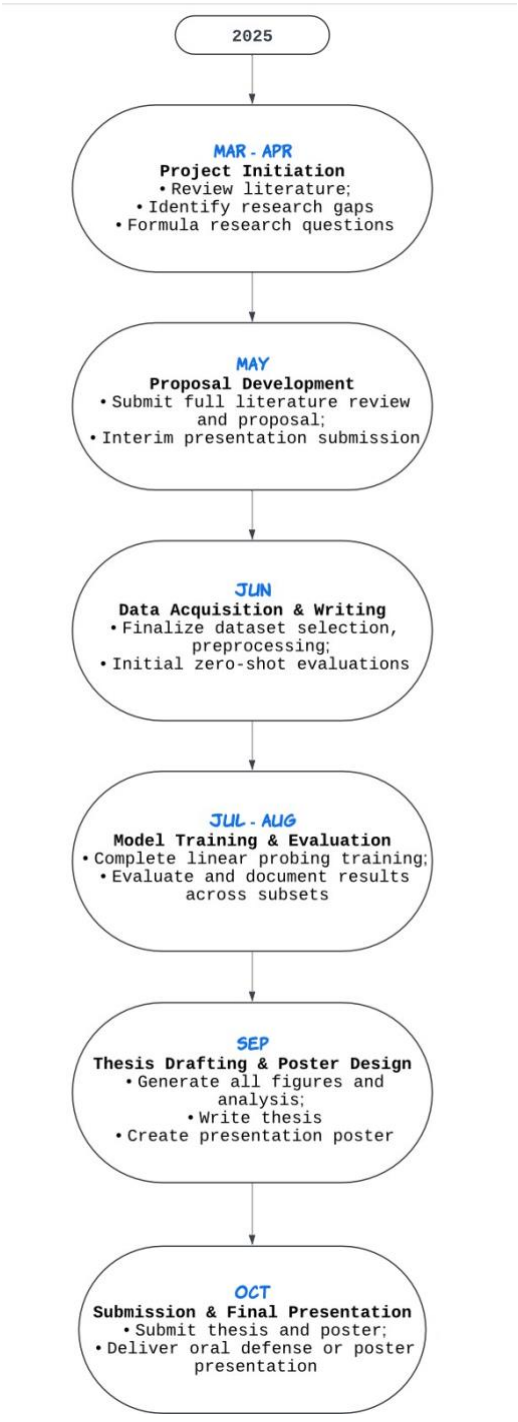


Figure 4

Research project timeline outlining key milestones across the 2025 academic calendar. Activities include literature review, proposal submission, dataset preparation, model training (zero-shot and linear probing), evaluation, and thesis/poster development.

4.5 Limitations and Risk Management

While the proposed methodology is grounded in best practices, several limitations must be acknowledged. First, despite the robust performance reported by Silva-Rodríguez et al. (2025), zero-shot inference may underperform in real-world clinical settings, especially in cases involving subtle pathologies or comorbid conditions. Second, the performance of vision-language models like FLAIR is sensitive to prompt formulation, with small variations in phrasing leading to significant differences in classification outcomes (Radford et al., 2021; Silva-Rodríguez et al., 2025). Third, public datasets such as Messidor and REFUGE, while widely adopted, may include label noise and lack sufficient demographic or device diversity to generalize well across populations (Zhou et al., 2023).

Additionally, FLAIR's dual-encoder architecture introduces considerable computational overhead, particularly during embedding alignment and repeated evaluation across multiple random seeds. This may limit experimentation in constrained computing environments. Finally, training linear probes on extremely small subsets for instance, 1%, poses risks of overfitting and unstable evaluation due to sample variance.

To manage these risks, this project incorporates multiple safeguards: using stratified 5-fold cross-validation to mitigate sampling bias; performing multiple repetitions with different random seeds; harmonizing preprocessing across all conditions; and publishing the entire codebase under an open-source license for full reproducibility. These limitations and mitigations are transparently documented to guide result interpretation and inform future iterations of foundation model deployment in medical imaging.

4.6 Ethics and Data Privacy Considerations

This project uses fully de-identified, publicly available datasets, Messidor and REFUGE, under open licenses, with no PII or human interaction involved (Decencière et al., 2014; Orlando et al., 2020). As per NHMRC (2023) and Monash guidelines (2024), ethics approval is not required. **Therefore, this project does not require any ACS Ethics approval.**

Conclusion

This study presents a systematic investigation into the adaptation performance of the FLAIR vision-language foundation model for retinal disease diagnosis, focusing on two prevalent conditions: diabetic retinopathy (DR) and glaucoma. By designing a controlled comparison between zero-shot inference and linear probing on the Messidor and REFUGE datasets, both excluded during FLAIR's pretraining, this research addresses a critical gap in current ophthalmic AI literature.

The proposed experimental framework simulates real-world, low-resource clinical environments through strategic data down sampling and rigorous evaluation protocols. This enables the assessment of adaptation strategies under conditions reflective of deployment in rural, underserved, or data-scarce settings. The study not only evaluates FLAIR's generalization through zero-shot inference but also measures performance gains from lightweight adaptation via linear probing.

This approach yields three key contributions: (1) it clarifies adaptation trade-offs between general-purpose inference and task-specific tuning in vision-language models for ophthalmology, (2) it introduces a reproducible benchmark for comparing adaptation strategies using standardized datasets and prompt formulations, and (3) it offers practical insights into deploying foundation models like FLAIR without extensive retraining.

In doing so, the study provides a reproducible evaluation framework that informs future adaptation strategies for medical AI in ophthalmology, particularly under low-data constraints. The findings are expected to inform the development of vision-language models for both common and rare ophthalmic diseases, particularly in settings with limited labelled data, computational resources,

or clinical oversight. As foundation models continue to transform medical imaging, this work offers a timely and rigorous framework for adapting and evaluating these systems responsibly and effectively.

References

Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *Npj Digital Medicine*, *1*(1), 39. https://doi.org/10.1038/s41746-018-0040-6

Ayhan, M. S., Ong, A. Y., Ruffell, E., Wagner, S. K., Merle, D. A., & Keane, P. A. (2025). In-context learning for data-efficient classification of diabetic retinopathy with multimodal foundation models. *MedRxiv*. https://doi.org/10.1101/2025.03.09.25323618

Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., Natarajan, V., & Norouzi, M. (2021). *Big Self-Supervised Models Advance Medical Image Classification(pp. 3478-3488)*. Openaccess.thecvf.com. https://openaccess.thecvf.com/content/ICCV2021/html/Azizi_Big_Self-Supervised_Models_Advance_Medical_Image_Classification_ICCV_2021_paper.html

Berkowitz, S. T., Finn, A. L., Parikh, R., Kuriyan, A. E., & Patel, S. (2023). Ophthalmology Workforce Projections in the United States, 2020-2035. *Ophthalmology*, *131*(2), 133–139. https://doi.org/10.1016/j.ophtha.2023.09.018

Bourne, R. R. A., Stevens, G. A., White, R. A., Smith, J. L., Flaxman, S. R., Price, H., Jonas, J. B., Keeffe, J., Leasher, J., Naidoo, K., Pesudovs, K., Resnikoff, S., & Taylor, H. R. (2013). Causes of vision loss worldwide, 1990–2010: a systematic analysis. *The Lancet Global Health*, *1*(6), e339–e349. https://doi.org/10.1016/s2214-109x(13)70113-x

Chang, Y., Jiang, Z., Zhang, K., & Zhou, S. K. (2025). *FundusAdapter: few-shot adaptation of fundus image foundation model for fundus image diagnosis*. OpenReview. https://openreview.net/forum?id=enWkUsJyjf#discussion

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., Erginay, A., Charton, B., & Klein, J.-C. (2014). FEEDBACK ON A PUBLICLY DISTRIBUTED IMAGE DATABASE: THE MESSIDOR DATABASE. *Image Analysis & Stereology*, *33*(3), 231–234. https://doi.org/10.5566/ias.1155

Der, V., & Hinton, G. (2008). Viualizing data using t-SNE. Journal of Machine Learning Research, 9(11), 2579–2605. https://www.researchgate.net/publication/228339739_Viualizing_data_using_t-SNE

Du, J., Guo, J., Zhang, W., Yang, S., Liu, H., Li, H., & Wang, N. (2024). RET-CLIP: A Retinal Image Foundation Model Pre-trained with Clinical Diagnostic Reports. *Lecture Notes in Computer Science*, 709–719. https://doi.org/10.1007/978-3-031-72390-2_66

Emese Sükei, Rumetshofer, E., Niklas Schmidinger, Mayr, A., Schmidt-Erfurth, U., Günter Klambauer, & Hrvoje Bogunović. (2024). Improving Clinical Predictions with Multi-Modal Pre-training in Retinal Imaging. In 2024 IEEE International Symposium on Biomedical Imaging, 1–5. https://doi.org/10.1109/isbi56570.2024.10635447

Gao, M., Jiang, H., Zhu, L., Jiang, Z., Geng, M., Ren, Q., & Lu, Y. (2023). Discriminative ensemble meta-learning with co-regularization for rare fundus diseases diagnosis. *Medical Image Analysis*, *89*, 102884. https://doi.org/10.1016/j.media.2023.102884

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., & Qiao, Y. (2023). CLIP-Adapter: Better Vision-Language Models with Feature Adapters. *International Journal of Computer Vision*, *132*(2), 581–595. https://doi.org/10.1007/s11263-023-01891-x

Gholami, S., Jannat, F.-E., Thompson, A. C., Ong, S. S. Y., Lim, J. I., Leng, T., Tabkhivayghan, H., & Alam, M. N. (2025). Distributed training of foundation models for ophthalmic diagnosis. Communications Engineering, 4(1), 6. https://doi.org/10.1038/s44172-025-00341-5

Goscinski, W., McIntosh, P., Ulrich Claus Felzmann, Maksimenko, A., Hall, C., Gureyev, T. E., Thompson, D. A., Janke, A. L., Galloway, G. J., Neil, Parnesh Raniga, Kaluza, O., Ng, A., Govinda Poudel, Barnes, D., Nguyen, T. D., C. Paul Bonnington, & Egan, G. F. (2014). The multi-modal Australian ScienceS Imaging and Visualization Environment (MASSIVE) high performance computing infrastructure: applications in neuroscience and neuroinformatics research. Frontiers in Neuroinformatics, 8, 30. https://doi.org/10.3389/fninf.2014.00030

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P. C., Mega, J. L., & Webster, D. R. (2016). Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, *316*(22), 2402–2410. https://doi.org/10.1001/jama.2016.17216

Holmberg, O. G., Köhler, N. D., Martins, T., Siedlecki, J., Herold, T., Keidel, L., Asani, B., Schiefelbein, J., Priglinger, S., Kortuem, K. U., & Theis, F. J. (2020). Self-supervised retinal thickness prediction enables deep learning from unlabelled data to boost

classification of diabetic retinopathy. *Nature Machine Intelligence*, *2*(11), 719–726. https://doi.org/10.1038/s42256-020-00247-1

Jin, R., Xu, Z., Zhong, Y., Yao, Q., Dou, Q., K, Z. S., & Li, X. (2024). FairMedFM: Fairness Benchmarking for Medical Imaging Foundation Models. *Advances in Neural Information Processing Systems*, *37*, 111318–111357. https://proceedings.neurips.cc/paper_files/paper/2024/hash/c9826b9ea5e1b49b256329934a578d83-Abstract-Datasets_and_Benchmarks_Track.html

Kumar, Y., & Marttinen, P. (2024). Improving Medical Multi-modal Contrastive Learning with Expert Annotations. *Lecture Notes in Computer Science*, 468–486. https://doi.org/10.1007/978-3-031-72661-3_27

Li, Y., Fu, Y., Yang, Q., Min, Z., Yan, W., Huisman, H., Barratt, D., Prisacariu, V. A., & Hu, Y. (2022). FEW-SHOT Image Segmentation for Cross-Institution Male Pelvic Organs Using Registration-Assisted Prototypical Learning. *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, *59*, 1–5. https://doi.org/10.1109/isbi52829.2022.9761453

Li, Z., Song, D., Yang, Z., Wang, D., Li, F., Zhang, X., Kinahan, P. E., & Qiao, Y. (2024). VisionUnite: A Vision-Language Foundation Model for Ophthalmology Enhanced with Clinical Knowledge. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2408.02865

Minutti, C. (2024). *Enhancing Interpretability and Fairness in Medical Foundation Models: A Generative Approach for Explainable and Bias-Mitigated Medical Image Analysis*. OpenReview. https://openreview.net/forum?id=T2wOaTwyk8

Murugappan, M., Prakash, N. B., Jeya, R., Mohanarathinam, A., Hemalakshmi, G. R., & Mahmud, M. (2022). A novel few-shot classification framework for diabetic retinopathy

detection and grading. *Measurement*, *200*, 111485. https://doi.org/10.1016/j.measurement.2022.111485

Nguyen, H., Nguyen, H., Diep, N., Pham, T. N., Cao, T., Nguyen, B., Swoboda, P., Ho, N., Albarqouni, S., Xie, P., Sonntag, D., & Niepert, M. (2023). LVM-Med: Learning Large-Scale Self-Supervised Vision Models for Medical Imaging via Second-order Graph Matching. *Advances in Neural Information Processing Systems*, *36*, 27922–27950. https://proceedings.neurips.cc/paper_files/paper/2023/hash/58cc11cda2a2679e8af5c6317aed0af8-Abstract-Conference.html

Oakden-Rayner, L., Dunnmon, J., Carneiro, G., & Re, C. (2020). Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *Proceedings of the ACM Conference on Health, Inference, and Learning*, 151–159. https://doi.org/10.1145/3368555.3384468

Orlando, J. I., Fu, H., Barbosa Breda, J., van Keer, K., Bathula, D. R., Diaz-Pinto, A., Fang, R., Heng, P.-A., Kim, J., Lee, J., Lee, J., Li, X., Liu, P., Lu, S., Murugesan, B., Naranjo, V., Phaye, S. S. R., Shankaranarayana, S. M., Sikka, A., & Son, J. (2020). REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical Image Analysis*, *59*, 101570. https://doi.org/10.1016/j.media.2019.101570

Pachetti, E., & Colantonio, S. (2024). A systematic review of few-shot learning in medical imaging. *Artificial Intelligence in Medicine*, *156*, 102949. https://doi.org/10.1016/j.artmed.2024.102949

Pan, L., Zhang, P., Xia, F., Ji, B., Liu, W., Wang, H., Jin, Y., Chongcheawchamnan, M., & Peng, S. (2021). FEDI: Few-shot learning based on Earth Mover's Distance algorithm

combined with deep residual network to identify diabetic retinopathy. *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1032–1036. https://doi.org/10.1109/bibm52615.2021.9669547

Park, I., Kim, S., & Ryu, J. (2024). Generative Self-Supervised Learning for Medical Image Classification. *Thecvf.com*, 976–993. https://openaccess.thecvf.com/content/ACCV2024/html/Park_Generative_Self-Supervised_Learning_for_Medical_Image_Classification_ACCV_2024_paper.html

Peng, Z., Ma, R., Zhang, Y., Yan, M., Lu, J., Qian, C., Liao, J., Zhang, Y., Wang, J., Zhao, Y., Zhu, J., Qin, B., Jiang, Q., Shi, F., Jiang, Q., Chen, X., & Zhao, C. (2023). Development and evaluation of multimodal AI for diagnosis and triage of ophthalmic diseases using ChatGPT and anterior segment images: protocol for a two-stage cross-sectional study. *Frontiers in Artificial Intelligence*, *6*, 1323924. https://doi.org/10.3389/frai.2023.1323924

Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., & Meriaudeau, F. (2018). Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research. *Data*, *3*(3), 25. https://doi.org/10.3390/data3030025

Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., Liu, L., Wang, J., Liu, X., Gao, L., Wu, T., Xiao, J., Wang, F., Yin, B., Wang, Y., Danala, G., He, L., Choi, Y. H., Lee, Y. C., & Jung, S.-H. (2020). IDRiD: Diabetic Retinopathy – Segmentation and Grading Challenge. *Medical Image Analysis*, *59*, 101561. https://doi.org/10.1016/j.media.2019.101561

Qian, B., Wang, X., Guan, Z., Yang, D., Ran, A., Li, T., Wang, Z., Wen, Y., Shu, X., Xie, J., Liu, S., Xing, G., Silva-Rodríguez, J., Riadh Kobbi, Li, P., Chen, T., Bi, L., Kim, J., Jia,

W., & Li, H. (2024). HRDC challenge: a public benchmark for hypertension and hypertensive retinopathy classification from fundus images. The Visual Computer, 41(2), 1061–1077. https://doi.org/10.1007/s00371-024-03384-5

Qiu, J., Wu, J., Wei, H., Shi, P., Zhang, M., Sun, Y., Li, L., Liu, H., Liu, H., Hou, S., Zhao, Y., Shi, X., Xian, J., Qu, X., Zhu, S., Pan, L., Chen, X., Zhang, X., Jiang, S., & Wang, K. (2023). VisionFM: a Multi-Modal Multi-Task Vision Foundation Model for Generalist Ophthalmic Artificial Intelligence. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2310.04992

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021, July 1). *Learning Transferable Visual Models From Natural Language Supervision(pp. 8748-8763))*. Proceedings.mlr.press; PMLR. https://proceedings.mlr.press/v139/radford21a

Rashidisabet, H., Sethi, A., Jindarak, P., Edmonds, J., Chan, R. V. P., Leiderman, Y. I., Vajaranant, T. S., & Yi, D. (2023). Validating the Generalizability of Ophthalmic Artificial Intelligence Models on Real-World Clinical Data. Translational Vision Science & Technology, 12(11), 231–234. https://doi.org/10.1167/tvst.12.11.8

Rishi Bommasani, Hudson, D. A., Ehsan Adeli, Altman, R. B., Arora, S., Sydney von Arx, Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N. S., Chen, A. T., Creel, K., Davis, J., Demszky, D., & Donahue, C. (2021). On the Opportunities and Risks of Foundation Models. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2108.07258

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017

IEEE International Conference on Computer Vision (ICCV), 618–626. https://doi.org/10.1109/iccv.2017.74

Sevgi, M., Ruffell, E., Antaki, F., Chia, M. A., & Keane, P. A. (2024). Foundation models in ophthalmology: opportunities and challenges. *Current Opinion in Ophthalmology*, *36*(1), 90–98. https://doi.org/10.1097/icu.0000000000001091

Shi, C., Rezai, R., Yang, J., Dou, Q., & Li, X. (2024). *A Survey on Trustworthiness in Foundation Models for Medical Image Analysis*. ArXiv.org. https://arxiv.org/abs/2407.15851

Shi, D., Zhang, W., Chen, X., Liu, Y., Yang, J., Huang, S., Tham, Y. C., Zheng, Y., & He, M. (2024). EyeFound: A Multimodal Generalist Foundation Model for Ophthalmic Imaging. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2405.11338

Sokolova, M., & Lapalme, G. (2009). A Systematic Analysis of Performance Measures for Classification Tasks. Information Processing & Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Silva-Rodríguez, J., Hadi Chakor, Riadh Kobbi, Dolz, J., & Ismail Ben Ayed. (2024). A Foundation Language-Image Model of the Retina (FLAIR): Encoding expert knowledge in text supervision. *Medical Image Analysis*, *99*, 103357. https://doi.org/10.1016/j.media.2024.103357

Tibshirani, R. J., & Efron, B. (1993). An introduction to the bootstrap. Monographs on statistics and applied probability, 57(1), 1-436. Chapman and Hall/CRC. https://doi.org/10.1201/9780429246593

Ting, D. S. W., Cheung, C. Y.-L., Lim, G., Tan, G. S. W., Quang, N. D., Gan, A., Hamzah, H., Garcia-Franco, R., San Yeo, I. Y., Lee, S. Y., Wong, E. Y. M., Sabanayagam, C.,

Baskaran, M., Ibrahim, F., Tan, N. C., Finkelstein, E. A., Lamoureux, E. L., Wong, I. Y., Bressler, N. M., & Sivaprasad, S. (2017). Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA*, *318*(22), 2211–2223. https://doi.org/10.1001/jama.2017.18152

Ting, D. S. W., Pasquale, L. R., Peng, L., Campbell, J. P., Lee, A. Y., Raman, R., Tan, G. S. W., Schmetterer, L., Keane, P. A., & Wong, T. Y. (2019). Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*, *103*(2), 167–175. https://doi.org/10.1136/bjophthalmol-2018-313173

Waskom, M. (2021). Seaborn: Statistical Data Visualization. Journal of Open Source Software, 6(60), 3021. https://doi.org/10.21105/joss.03021

Wang, M., Gong, Q., Wan, Q., Leng, Z., Xu, Y., Yan, B., Zhang, H., Huang, H., & Sun, S. (2023). A fast interpretable adaptive meta-learning enhanced deep learning framework for diagnosis of diabetic retinopathy. *Expert Systems with Applications*, *244*, 123074. https://doi.org/10.1016/j.eswa.2023.123074

Wang, M., Lin, T., Yu, K., Lin, A., Peng, Y., Wang, L., Chen, C., Zou, K., Liang, H., Chen, M., Yao, X., Zhang, M., Huang, B., Zheng, C., Chen, W., Luo, Y., Chen, Y., Wang, J., Tham, Y. C., & Liu, D. (2024). Common and Rare Fundus Diseases Identification Using Vision-Language Foundation Model with Knowledge of Over 400 Diseases. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2406.09317

Wang, S., He, X., Jian, Z., Li, J., Xu, C., Chen, Y., Liu, Y., Chen, H., Huang, C., Hu, J., & Liu, Z. (2024). Advances and prospects of multi-modal ophthalmic artificial intelligence

based on deep learning: a review. *Eye and Vision*, *11*(1), 38. https://doi.org/10.1186/s40662-024-00405-1

World Health Organization. (2023, August 10). *Blindness and vision impairment*. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment

Wu, R., Su, N., Zhang, C., Ma, T., Zhou, T., Cui, Z., Tang, N., Mao, T., Zhou, Y., Fan, W., Wu, T., Jing, S., & Fu, H. (2025). MM-Retinal V2: Transfer an Elite Knowledge Spark into Fundus Vision-Language Pretraining. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2501.15798

Wu, R., Zhang, C., Zhang, J., Zhou, Y., Zhou, T., & Fu, H. (2024). MM-Retinal: Knowledge-Enhanced Foundational Pretraining with Fundus Image-Text Expertise. *Lecture Notes in Computer Science*, 722–732. https://doi.org/10.1007/978-3-031-72378-0_67

Yang, K. O., Lee, J. M., Shin, Y., Yoon, I. Y., Choi, J. W., & Lee, W. J. (2024). Diagnosis of Glaucoma Based on Few-Shot Learning with Wide-Field Optical Coherence Tomography Angiography. *Biomedicines*, *12*(4), 741. https://doi.org/10.3390/biomedicines12040741

Yang, S., Du, J., Guo, J., Zhang, W., Liu, H., Li, H., & Wang, N. (2024). ViLReF: A Chinese Vision-Language Retinal Foundation Model. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2408.10894

Yu, K., Zhou, Y., Bai, Y., Soh, Z. D., Xu, X., Goh, R. S. M., Cheng, C.-Y., & Liu, Y. (2024). UrFound: Towards Universal Retinal Foundation Models via Knowledge-Guided Masked Modeling. *Lecture Notes in Computer Science*, 753–762. https://doi.org/10.1007/978-3-031-72390-2_70

Zhou, Y., Chia, M. A., Wagner, S. K., Ayhan, M. S., Williamson, D. J., Struyven, R. R., Liu, T., Xu, M., Lozano, M. G., Woodward-Court, P., Kihara, Y., Altmann, A., Lee, A. Y., Topol, E. J., Denniston, A. K., Alexander, D. C., & Keane, P. A. (2023). A foundation model for generalizable disease detection from retinal images. *Nature*, *622*(7981), 156–163. https://doi.org/10.1038/s41586-023-06555-x

Zou, K., Chen, Z., Yuan, X., Shen, X., Wang, M., & Fu, H. (2023). A review of uncertainty estimation and its application in medical imaging. *Meta-Radiology*, *1*(1), 100003. https://doi.org/10.1016/j.metrad.2023.100003

Part II

Main Thesis

FROM PROMPTS TO PROBES: ZERO-SHOT AND FEW-SHOT TRANSFER OF FOUNDATION MODELS FOR RETINAL FUNDUS CLASSIFICATION

FROM PROMPTS TO PROBES: ZERO-SHOTAND FEW-SHOT TRANSFER OF FOUNDATION MODELS FOR RETINAL FUNDUS CLASSIFICATION

Pugalenthi Magendran[1], Yasmeen George[1]*

Faculty of Information Technology, Monash University, Melbourne, Australia

ABSTRACT

Screening and triage from retinal fundus photographs increasingly rely on foundation models, yet their behaviour across tasks and under scarce labels remains poorly characterised. We compare four encoders across three benchmarks. OpenCLIP serves as a generalist model, BiomedCLIP as a biomedical vision and language model, FLAIR as a retina tuned vision and language model, and RETFound as a retina self-supervised model. The benchmarks are MESSIDOR for diabetic retinopathy, ODIR 200×3 for multiple diseases, and REFUGE for glaucoma. We assess two adaptation regimes. The first is zero-shot classification with clinically phrased prompts. The second is few-shot linear probing with 5%, 10%, and 20% labels. Evaluation is decision centred and covers Macro-F1 on MESSIDOR and ODIR-200×3, AUROC and sensitivity at 90% specificity on REFUGE, and calibration using ECE and AUCE, as well as prompt sensitivity.

Across datasets, FLAIR is the most consistent zero-shot performer on MESSIDOR and REFUGE, while BiomedCLIP leads on ODIR 200×3. With small label budgets, linear probes narrow the gaps and, on REFUGE, can reverse the ranking in favour of OpenCLIP. Simple temperature scaling improves calibration without altering discrimination, and prompt choice measurably affects zero-shot results. The study standardises a prompt to probes protocol, aligns metrics with clinical endpoints and fixed operating points, and offers practical guidance on when prompts suffice and when a lightweight probe is warranted for safe and reproducible deployment in retinal screening.

1. INTRODUCTION

Automated analysis of retinal fundus photographs supports screening and management for conditions such as diabetic retinopathy (DR) and glaucoma, where timely detection can prevent vision loss. Global projections estimate that the number of adults with DR will increase from approximately 103 million in 2020 to approximately 160 million by 2045 [1]. This sustained burden underscores the need for scalable approaches to screening and triage.

MESSIDOR contains 1,200 colour fundus photographs of the posterior pole with a 45° field of view, collected at three sites with varying resolutions. Of these, 800 images are dilated and 400 are non-dilated [2]. The dataset provides an image level medical diagnosis without lesion annotations, which supports evaluation of robustness under realistic acquisition variability and comparative assessment without per lesion labels [2].

REFUGE provides a rigorously controlled benchmark for glaucoma assessment with 1,200 colour fundus photographs and fixed training, offline test, and online test partitions, which enables uniform comparison across methods [3]. For glaucoma classification, the REFUGE benchmark reports AUROC and the sensitivity at a fixed 85% specificity, which supports operating point analysis alongside rank order metrics. For structure prediction, optic disc and cup segmentation is evaluated with Dice and the mean absolute error of the vertical cup to disc ratio [3].

OIA ODIR captures real world heterogeneity: 10,000 fundus images from 5,000 patients collected on different cameras across multiple ophthalmic centres, with three predefined splits (training, off site test, on site test) and eight clinical categories (N, D, G, C, A, H, M, O) [4]. The dataset shows pronounced class imbalance and variation across devices and sites, making it a natural stress test for recognition across multiple diseases [4]. In this study we use the ODIR 200×3 subset (Normal, Cataract, Pathologic Myopia) to enable a controlled and balanced evaluation, following prior work [5, 9].

Over recent years, foundation models have reshaped the design space for retinal AI. Vision-language encoders, exemplified by CLIP, jointly train an image encoder and a text encoder with a contrastive objective so that paired images and texts are close in a shared space. They perform zero-shot classification by embedding class names or descriptions and choosing the label with the highest similarity to the image representation [6]. In practice, prompt choice matters. CLIP style evaluations often ensemble multiple prompt templates per class, and template phrasing measurably affects zero-shot accuracy [6, 7].

At one end of the spectrum, generalist CLIP variants trained on internet image and text pairs, such as OpenCLIP, provide competitive and reproducible baselines with open-source code, released models, and well documented training and evaluation pipelines that describe CLIP scaling laws [7].

Domain tuned vision-language models aim to close the semantic gap. BiomedCLIP is pre trained on PMC 15M, which contains approximately 15 million biomedical figure and caption pairs from PubMed Central, using contrastive learning, and it transfers strongly to retrieval, zero-shot image classification, and visual question answering [8]. FLAIR is a retina specific VLMs that aligns fundus images with expert clinical text such as lesion level descriptors, learning encoders suited to downstream retinal tasks [9]. In parallel, self-supervised learning in the retina domain is typified by RETFound, which uses masked autoencoder pre training on large unlabelled ocular image corpora covering fundus photographs and OCT to learn anatomy and lesion patterns efficiently, then adapts to clinical tasks with modest labelled data through fine tuning [10].

Clinical and operational constraints. Retinal AI is often deployed where labelled data are scarce and disease prevalence at the point of screening is low, while site and domain shift is unavoidable because cameras, populations, and acquisition protocols differ between pretraining sources and the target clinic [1,4].

Under these constraints, two transfer strategies are especially practical: zero-shot vision-language inference, which requires no local tuning and uses clinician faithful class prompts, and few-shot linear probes on frozen encoders, a lightweight adaptation when small, labelled sets can be obtained via retrospective chart audit or a brief labelling sprint, keeping compute and governance overhead low.

A like for like comparison of these regimes across DR grading, glaucoma screening, and multi disease recognition can be conducted on MESSIDOR [2], REFUGE [3], and ODIR 200×3, a balanced three class subset of OIA ODIR [4], directly addressing a real deployment decision point for ophthalmology services.

Despite the momentum, two practical questions remain underexplored for retinal applications:

(i) How do vision and language (VL) and self-supervised learning (SSL) encoders transfer across heterogeneous retinal tasks when supervision is scarce? Many clinical deployments begin with none or very few labels. Zero-shot approaches promise no tuning utility via prompts, and few shot linear probes promise rapid adaptation without full fine tuning [6 to 10]. To our knowledge, systematic and controlled comparisons across DR grading (MESSIDOR [2]), glaucoma screening (REFUGE [3]), and multi-disease recognition (ODIR 200 x 3 [4]) within a single protocol remain limited.

(ii) How robust are prompt based zero-shot predictions to class phrasing, and how calibrated are probabilities once a small linear head is trained? Zero-shot CLIP constructs a classifier by embedding class names and descriptions as text prompts, and performance varies with prompt templates and prompt ensembling [6] (see also follow ups on scaling and evaluation practice [7]). When limited labels are available, CLIP commonly fits a simple linear classifier on frozen features, which motivates explicit evaluation of probability calibration after such adaptation [6]. Poor calibration can undermine triage decisions even when discrimination is strong, and post hoc temperature scaling is an effective baseline. Recent work further highlights calibration as essential in safety critical settings [11 to 12]. In ophthalmic screening, where clinical action is threshold based, operating point behaviour and calibration are as important as rank ordering.

This study addresses these gaps through a controlled evaluation across multiple datasets that covers prompt based zero-shot inference and few shot linear probing. We compare a general-purpose CLIP baseline (OpenCLIP), a biomedical vision-language encoder (BiomedCLIP), a retina tuned vision language model (VLM), and a retinal self-supervised encoder (RETFound) on three representative benchmarks: MESSIDOR for diabetic retinopathy grading, ODIR 200×3 for three class fundus disease classification, and REFUGE for glaucoma screening [2–4,6–10]. We examine two complementary transfer regimes. Zero-shot uses CLIP style cosine similarity between image embeddings and prompted class embeddings, creating a text conditioned classifier without any local tuning.

To quantify prompt sensitivity, we evaluate three clinically valid templates per class (T1–T3) that vary lexical phrasing and the inclusion of salient descriptors such as DR lesions, optic nerve head cues for glaucoma, and disease names for ODIR [6,8–9]. Few shot linear probing freezes the image encoder and trains a single linear classifier to measure representation quality without the confounds of end-to-end fine tuning. We use small, labelled subsets at 5%, 10%, and 20%, consistent with prior work on label efficiency [6,9].

Datasets and task aware endpoints. MESSIDOR comprises 1,200 posterior pole colour fundus photographs from three sites, including 800 dilated and 400 non dilated eyes, captured with a 45° field of view. It provides image level diagnoses suitable for DR grading, and we summarise performance with macro averaged or class balanced metrics [2]. ODIR 200×3 is a balanced three class subset of ODIR 5K comprising Normal, Cataract, and Pathologic Myopia, with 200 images per class.

We use it for controlled class balanced evaluation and report Macro F1 as the primary endpoint, with Average Class Accuracy (ACA) as a secondary metric for comparability with prior work [4,9]. REFUGE is a rigorously controlled glaucoma benchmark with 1,200 colour fundus images and fixed training, offline test, and online test partitions. Official reporting includes AUROC and a reference sensitivity at 85% specificity, supporting operating point analyses alongside rank order metrics [3]. Primary endpoints are Macro F1 for MESSIDOR and ODIR 200×3 and AUROC for REFUGE, complemented by operating point and calibration analyses where appropriate.

Why operating points and calibration matter. When disease prevalence is low and the false positive load affects operations, AUROC alone is not sufficient. We therefore report metrics at clinically relevant operating points and examine precision and recall behaviour in imbalanced settings [15].

We also assess probability calibration using expected calibration error, both with and without post hoc temperature scaling. Temperature scaling is a single parameter rescaling of logits fitted on the validation set that improves the quality of confidence estimates while leaving class rankings and accuracy unchanged, and therefore does not change AUROC. This is important for threshold-based screening decisions [11, 12].

Risk framing: shortcuts and hidden stratification. Medical imaging datasets often contain spurious site or device cues that models exploit as shortcuts, degrading performance under distribution shift. Unlabelled subgroups can produce hidden stratification, yielding clinically meaningful failure modes even when global AUROC is strong.

These risks motivate two practices. First, report prompt sensitivity by quantifying variance across clinically valid prompts to surface fragility in zero shot semantics. Second, emphasise calibration and the selection of operating points, because brittle probability estimates or unstable thresholds are early warnings even when average discrimination is adequate [13, 14].

Contributions:
(1) Unified prompts to probes protocol for retinal transfer. We standardise zero shot prompt evaluation using three clinician validated templates per class, and few shot linear probes with 5 percent, 10 percent, and 20 percent labels across heterogeneous tasks: MESSIDOR, ODIR 200×3, and REFUGE. We report per prompt results, prompt sensitivity summaries, and 95 percent confidence intervals.

(2) Head-to-head encoder comparison under identical conditions. We evaluate a general-purpose CLIP baseline OpenCLIP, a biomedical VLM BiomedCLIP, a retina tuned vision language model FLAIR, and a retina self-supervised encoder RETFound. This assesses whether promptable VLMs excel without labels and whether simple linear probes can narrow, or invert zero shot gaps once modest labels are available [6 to 10].

(3) Prompt sensitivity and calibration as first-class signals. We quantify variance across clinically valid prompts and assess calibration using ECE before and after temperature scaling, highlighting cases where confidence quality lags discrimination, which is critical for screening workflows [6, 11, 12].

(4) Task aware endpoints and operating point analysis. For MESSIDOR, we use image level grading metrics. For REFUGE, we include operating point analysis alongside AUROC, reflecting the benchmark's fixed specificity reporting. For ODIR 200×3, we report class balanced metrics consistent with prior work [2, 3, 4].

Overview:
Section 2 reviews related work on retinal foundation and This study compared zero-shot prompts and few-shot linear probes for retinal fundus classification across MESSIDOR, ODIR-200×3, and REFUGE using one consistent evaluation setup. A clear pattern emerges. FLAIR is the safest choice for diabetic-retinopathy grading on MESSIDOR in both zero-shot and few-shot settings. On ODIR-200×3, BiomedCLIP leads when labels are scarcest at five to ten percent, but FLAIR moves ahead once the label budget reaches twenty percent. For glaucoma on REFUGE, OpenCLIP overtakes the others when a simple linear probe is allowed, even though FLAIR is strongest in pure zero-shot. These trends hold across folds, seeds, and prompt templates.

The practical guidance is straightforward. If you must start without labels, FLAIR provides a solid zero-shot option for DR triage. If you can afford a small amount of annotation, a frozen-encoder linear probe is highly label-efficient on DR and ODIR-200×3 and often closes the gap to, or surpasses, zero-shot. For glaucoma screening where area under the ROC curve and sensitivity at a fixed specificity matter most, OpenCLIP with a linear probe is a strong default. Alongside rank-order metrics, reporting sensitivity at 90% specificity and calibration measures gives a more faithful picture of clinical utility. Simple temperature scaling improves confidence calibration while leaving AUROC unchanged.

This work has limits. We focus on three datasets, a small prompt set, and frozen-feature probes. Future extensions are direct: broaden disease and device coverage, expand the prompt catalogue, and test lightweight adapters and richer calibration. The unified prompts-to-probes protocol we used makes these additions easy and keeps results comparable.

Overall, the evidence supports selecting models by task and label budget rather than expecting a single winner in all settings.

- Section 2 reviews related work on retinal foundation/VL models and calibration.
- Section 3 presents datasets, models, prompt design, and evaluation protocols.
- Section 4 reports zero shot and few shot results with analyses of prompt sensitivity and calibration.
- Section 5 considers implications for deployment under label scarcity.
- Section 6 summarizes ethics, governance, and data privacy.
- Section 7 describes threats to validity and reproducibility.
- Section 8 concludes.

Aim:
To provide a decision ready comparison of zero shot prompts and few shot probes for diabetic retinopathy grading, glaucoma screening, and multi disease recognition under label scarcity.

2. BACKGROUND AND RELATED WORK

2.1 Retinal benchmarks and clinical problem framing

Automated analysis of retinal fundus photographs supports screening and longitudinal management in ophthalmology. The global burden of diabetic retinopathy remains substantial and is projected to increase from 103.12 million in 2020 to 160.50 million by 2045 [1].

Public benchmarks have shaped how progress is measured. MESSIDOR provides 1,200 posterior pole colour fundus images with per image clinical diagnoses of diabetic retinopathy, captured across three sites with realistic acquisition variability, including dilation status and resolution. It does not include per lesion annotations, which facilitates comparative evaluation [2].

REFUGE offers a unified benchmark for glaucoma with fixed training, offline test, and online test partitions, along with stratified prevalence and standardised evaluation. This design supports operating point analysis alongside AUROC for screening use cases [3].

ODIR aggregates heterogeneous fundus images across multiple ocular conditions, including diabetic retinopathy, glaucoma, myopia, and age-related macular degeneration. It reflects clinic like label diversity and acquisition variability across centres and devices, enabling recognition of multiple diseases [4].

These datasets complement one another: REFUGE emphasises binary screening and safe triage, while MESSIDOR and ODIR emphasise multi class performance and robustness to label imbalance and taxonomy mismatch.

2.2 General vision-language pretraining and scaling

Open vocabulary recognition commonly follows the CLIP paradigm. A visual encoder is paired with a text encoder and trained with a contrastive objective on very large image and text corpora [6]. Classification is recast as comparing an image embedding with text conditioned class prototypes, which enables zero-shot transfer without task specific tuning. Empirically, contrastive language and image models exhibit scaling laws. Increasing compute, data, and capacity yields predictable gains across downstream tasks. This pattern motivates careful baseline selection and transparent reporting of size and data differences in comparisons [7].

EVA CLIP improves zero-shot transfer by combining stronger EVA initialized vision backbones with refined training techniques, achieving higher zero-shot performance than prior CLIP variants [16]. SigLIP 2 refines image and text matching by using a sigmoid loss together with auxiliary decoder based and self-supervised objectives. It improves semantic understanding, strengthens localization, produces higher quality dense features, and broadens multilingual coverage [17]. Hybrid objectives that combine image self-supervision with language aligned pretraining, exemplified by SLIP, improve representation quality in limited label settings and yield consistent gains in zero-shot transfer and linear probe evaluation [18]. TinyCLIP uses cross modal distillation with affinity mimicking, weight inheritance, and progressive stages to compress CLIP efficiently. It maintains competitive zero-shot accuracy while reducing parameters and training cost, which enables deployment in resource constrained environments [19].

Taken together, CLIP-style models form an evolving family in which the backbone, tokenizer, and training data materially affect medical transfer. Controlled studies should therefore isolate adaptation strategy, prompts versus probes, from backbone capacity.

2.3 Biomedical and retina-specialised foundation models

Domain aligned pretraining narrows the gap between web imagery and clinical data. BiomedCLIP aligns biomedical figures with captions at scale using PMC 15M, approximately fifteen million pairs, to capture medical image text semantics that generic web data miss. It provides a strong biomedical vision-language baseline for zero and few shot evaluation [8].

FLAIR is a retina specific vision-language model trained on thirty-eight public fundus datasets spanning one hundred and one target categories. It injects expert descriptors and

hierarchical clinical knowledge into the text encoder during pretraining and during zero-shot inference, with the goal of robust zero-shot fundus classification and generalisation under domain shifts [9].

In parallel, RETFound represents the self-supervised alternative. It is a masked autoencoder pretrained on about one point six million unlabelled retinal images, including fundus and OCT, to learn anatomy and lesion priors that can be adapted efficiently with lightweight heads [10].

Early follow ups test the transferability of retinal foundation models to less common conditions such as hypertensive retinopathy and to multicentre settings, documenting both strengths and limitations for generalisation [20].

Recent multi benchmark work, exemplified by MM Retinal V2, encourages broader cross benchmark evaluations by providing a public multimodality dataset and by reporting results under consistent zero-shot, few shot, and linear probing settings [21].

2.4 Adapting pre-trained encoders: zero-shot prompts, linear probes, and parameter-efficient tuning

Zero-shot prompting computes the cosine similarity between the image embedding and the text embeddings of candidate class prompts, and then predicts the class with the highest score without any task specific tuning [6].

When labels are limited, a common approach is to freeze the encoder and train a single linear classifier on the fixed features. This isolates representation quality and reduces overfitting relative to full fine tuning [20]. Calibrating the resulting predictions with a simple linear method can further improve accuracy and reliability in low label settings by reducing the expected calibration error and requiring only a small number of additional labelled examples [22].

Beyond these two endpoints, a substantial body of work examines parameter efficient adaptation for vision language models. CLIP Adapter augments a frozen CLIP with lightweight residual adapters placed after the encoders. This design blends pretrained features with features learned from few shot data and improves few shots performance and robustness under distribution shift [23]. TIP Adapter adopts a training free approach by caching few shot features in a key value store, delivering strong few shots performance with near zero training cost and minimal added inference compute [24].

On the text side, CoOp replaces discrete prompts with learnable continuous context, and CoCoOp conditions prompts on each input to reduce bias toward seen classes and improve generalization to unseen classes [25], [26]. Visual Prompt Tuning adapts frozen Vision Transformers by prepending learnable tokens and attains competitive performance with minimal parameter updates [27]. Broader studies of visual prompts confirm that input space prompting is a viable alternative when model weights cannot be modified [28].

Test time Prompt Tuning moves adaptation into the inference loop by optimizing prompt tokens for a single test sample, improving zero shot generalization without any downstream training data or annotations [29]. Prompt distribution learning models a distribution over prompts rather than a single template, producing diverse soft prompts that better match varying visual representations and improve few shots generalization [30]. Prompt aligned Gradient ProGrad stabilizes prompt tuning by aligning the update direction with zero-shot CLIP general knowledge, which reduces overfitting and brittleness under distribution shift [31].

Taken together, these methods motivate evaluation beyond headline zero-shot and few shot metrics. Studies should assess prompt robustness and adaptation stability, particularly in safety critical screening. Reports should include sensitivity and specificity at clinically relevant operating points, variation across random prompt seeds and templates, and stress tests under device, site, and population shift. They should also examine calibration, abstention or triage behaviour, and failure modes, and provide a clear accounting of model and data scale so that comparisons are fair and reproducible.

2.5 Medical vision–language alignment and low-data transfer

Taken together, these methods motivate evaluation that goes beyond headline zero-shot and few shot metrics. Studies should assess prompt robustness and adaptation stability, particularly in safety critical screening contexts.

Reports should present sensitivity and specificity at clinically relevant operating points, quantify variation across random prompt seeds and templates, and include stress tests under device, site, and population shift. They should also examine calibration, abstention or triage behaviour, and failure modes, and provide a clear accounting of model.

2.6 Metrics, operating points, and calibration in clinical screening

Metric choice must match task properties. For imbalanced problems, precision recall analysis is often more informative than ROC because ROC can conceal poor performance on the minority class [15]. Screening evaluations, exemplified by REFUGE, typically report operating point metrics such as sensitivity at a fixed specificity (85%) together with ROC AUC. We therefore also report Sens@Spec=90%, operating thresholds are selected on the validation split and applied unchanged to the test split. This practice aligns evaluation with triage settings that require low false positive rates [3].

ROC analysis has a long methodological history, including parametric and nonparametric approaches, and careful estimation is important when sample sizes are modest, and classes are skewed [39].

Decision support benefits from well calibrated probabilities. Modern neural networks are often overconfident, and temperature scaling provides a simple and effective post hoc recalibration. Calibration should be measured and reported together with accuracy. For example, Expected Calibration Error can be reported and complemented by the likelihood or the Brier score and by reliability diagrams [11-12].

Under dataset shift, both accuracy and the calibration of predictive uncertainty typically degrade, so it is advisable to evaluate uncertainty under shift rather than only on independent and identically distributed data [38]. From a statistical perspective, resampling methods such as the bootstrap provide confidence intervals and hypothesis tests. Comparing models on the same resamples or folds through paired analyses yields efficient uncertainty estimates. These tools provide principled uncertainty quantification when labels are limited [40].

2.7 Prompt robustness, dataset pathologies, and failure modes

Benchmarks:
We cover glaucoma screening with REFUGE, which reports sensitivity at fixed specificity along with ROC AUC. For multi disease and diabetic retinopathy recognition, we use ODIR and MESSIDOR. These datasets highlight different evaluation goals and practical considerations for deployment [2 to 4].

Backbones:
We compare four families. First, a general-purpose CLIP style baseline whose performance improves with greater data, compute, and model scale under language and image pretraining [6, 7, 16 to 19]. Second, a biomedical vision and language model with explicit alignment to clinical text, BiomedCLIP [8]. Third, a retina focused vision and language model that encodes expert descriptors, FLAIR [9]. Finally, a retina self-supervised encoder designed for label efficiency, RETFound [10].

Adaptation regimes:
We treat zero-shot prompting and few shot linear probing as a continuum and report results alongside parameter efficient alternatives that keep most model weights fixed, including adapters, prompt learning, visual prompts, and test time tuning. Examples include CLIP Adapter and Tip Adapter, CoOp and CoCoOp, Visual Prompt Tuning and visual prompting, Test time Prompt Tuning, and methods that stabilize or diversify prompt learning such as Prompt Distribution Learning and Prompt Aligned Gradient. These approaches emphasize strong transfer with small trainable parameter budgets or training free updates [23–31].

Evaluation:
We match metrics to tasks. For imbalanced classification we emphasize precision recall analysis since precision recall is more informative than ROC under skew [15]. For screening we report AUROC and sensitivity at a fixed specificity, following REFUGE [3].

We assess calibration and apply temperature scaling, reporting expected calibration error [11, 12]. We quantify prompt sensitivity as the mean and standard deviation across clinically valid templates. We quantify uncertainty with bootstrap confidence intervals and paired comparisons, including DeLong tests for AUC [3, 40].

This design differs from prior work in three ways. First, many studies evaluate a single vision-language or self-supervised family on a single retinal task. Second, they focus on tuned models while omitting zero shot baselines. Third, they report aggregate results under a single prompt template without sensitivity analysis. By standardising a prompt to probes protocol across heterogeneous retinal tasks, we address comparability gaps and foreground decision centric measures such as operating points and calibration that matter for safe deployment.

2.8 Governance and reporting

Although this section focuses on methods, clinical translation ultimately depends on disciplined reporting and sound governance. The CONSORT AI and SPIRIT AI extensions list AI specific items for trials, including input data acquisition and selection, algorithm versioning, human and AI interaction, and analysis of error cases. These checklists act as practical guardrails even during protocol development, helping teams design evaluations that stakeholders can trust [41, 42].

The WHO guidance on ethics and governance of AI for health emphasises transparency and explain ability, rigorous validation in the intended use settings with attention to subgroup performance, and life cycle risk management. These priorities align with reporting calibration, checking prompt sensitivity, and analysing uncertainty [43].

3. METHODOLOGY

3.1. Overviews

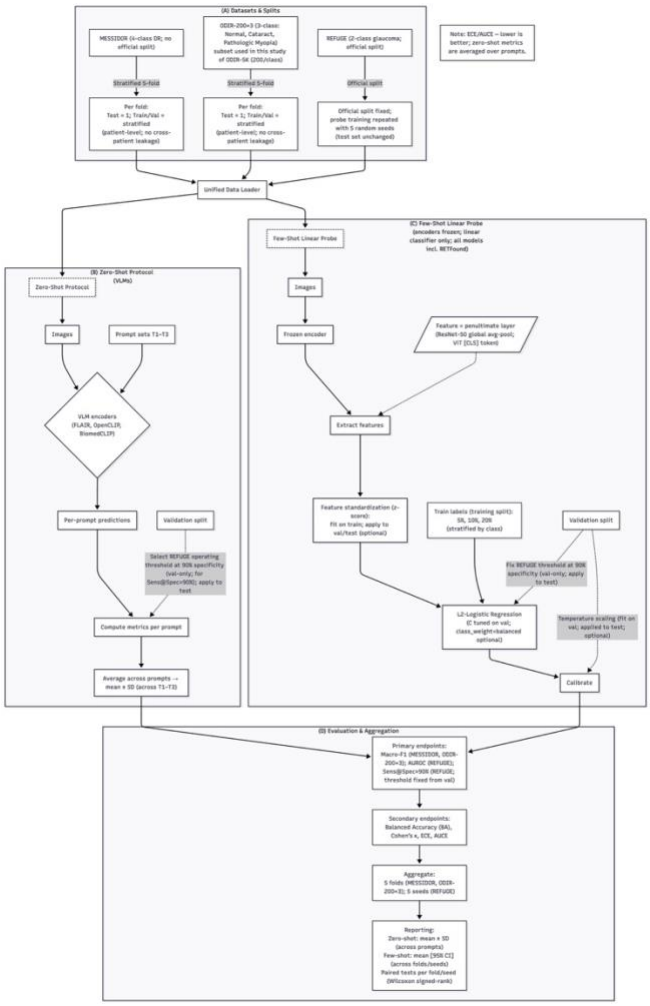


Fig 1. Summarizes the end-to-end pipeline for zero-shot and frozen-encoder linear-probe evaluations, with split policy, calibration, and reporting.

We evaluate vision and language encoders on retinal fundus classification under two protocols. The first uses zero-shot inference with multiple clinicians validated prompt templates. The second uses few shots linear probing with frozen encoders. We draw on three datasets that cover DR screening, multi disease recognition, and glaucoma assessment. These are MESSIDOR with four class DR [2], ODIR 200×3 with Normal, Cataract, and Pathologic Myopia which is a class balanced subset of ODIR 5K [4], and REFUGE with two class glaucoma [3].

REFUGE follows its official split. MESSIDOR and ODIR 200×3 use stratified fivefold splits. The primary endpoints are Macro F1 for MESSIDOR and ODIR 200×3 and AUROC for REFUGE. We also report Sens@Spec = 90% for REFUGE and secondary endpoints including Balanced Accuracy, Cohen's κ, ECE, and AUCE. Aggregation and significance testing are predefined and described in Sections 3.6 to 3.8.

Pre-processing: We load all images as RGB and apply the reference transform bundled with each checkpoint, including resize and crop steps and normalization by the mean and standard deviation used by that model. We do not use any augmentation during training because the encoders are frozen. For non-square inputs, the model specific transform pads or crops images as defined by the checkpoint implementation. Feature tensors are cached for each fold to prevent leakage from validation or test splits into training.

3.2. Datasets and splits

Dataset	Clinical Task	# Classes	Primary Endpoint	Train/Test Split
MESSIDOR	Diabetic retinopathy grading	4	Macro F1-Score	Stratified 5-fold; patient-level where IDs available (no official split)
REFUGE	Glaucoma Detection (glaucoma vs. non-glaucoma)	2	AUROC	Official fixed split (challenge): 400 train/ 400 val/ 400 test
ODIR-200×3, (subset of ODIR-5K)	Multi-disease recognition (Normal, Cataract, Pathologic Myopia)	3	Macro F1-Score	Custom subset (200/class); Stratified 5-fold (no official split)

Tab 1. Datasets and primary endpoints used in this study. Sources: MESSIDOR [2], REFUGE [3], ODIR-5K/ODIR-200×3 [4].

MESSIDOR (DR): Public benchmark for diabetic retinopathy with 1,200 colour fundus photographs drawn from three clinical sites and multiple resolutions. Each image includes a diagnostic label [2].

ODIR 200×3 (multi-disease): A balanced three class subset with Normal, Cataract, and Pathologic Myopia, derived from ODIR 5K which contains binocular pairs and eight disease categories [4]. We follow the FLAIR evaluation protocol [9].

REFUGE (glaucoma): Challenge dataset with fixed training, validation, and test partitions. Test labels are not publicly released. We use the official split [3].

Splitting:

(1) MESSIDOR and ODIR 200×3: We use stratified fivefold cross validation by class. When patient identifiers are available, fold assignment is at the patient level to avoid any overlap across patients.

(2) REFUGE: We use the official split. In the few shots protocol, only the linear probe is trained with five random seeds, and the test set remains unchanged. See Section 3.5.

Leakage policy: All thresholds and calibration parameters are selected on validation data only, then applied unchanged to the test set. Features are cached within each fold. We never precompute features using validation or test images when fitting training folds.

3.3. Models (Encoder)

Model Name	Vision Backbone	Text Backbone	Pretraining Data	Pretraining Data Size	Pretraining Objective
FLAIR	ResNet-50	BioClinicalBERT	38 retinal datasets	~288k images	Contrastive (expert prompts)
OpenCLIP	ViT-H/14	CLIP text transformer	LAION-2B (web captions)	2.32B image-text pairs	Contrastive language image (CLIP/InfoNCE)
BiomedCLIP	ViT-Base/16	PubMedBERT	PMC-15M	15M image-text pairs	Contrastive (scientific captions)
RETFound	ViT-Large	N/A	Retinal images	1.6M images	Masked Autoencoder (self-supervised)

Tab 2. Model backbones and pre-training sources used in this study. FLAIR [9], OpenCLIP/CLIP [6,7], BiomedCLIP [8], and RETFound [10].

FLAIR: retina specialized vision language foundation model that aligns fundus images with clinical text. We use the official checkpoints for zero shot evaluation and as frozen backbones for linear probing [9].

OpenCLIP: open reproduction of CLIP trained on large scale LAION datasets. We use it for zero-shot evaluation and linear probing [6,7].

BiomedCLIP: biomedical vision language model pretrained on approximately fifteen million scientific image text pairs. We use it for zero shot evaluation and linear probing [8].

RETFound: retina foundation model trained with self-supervision using a masked autoencoder on large retinal corpora. We use it only in the few shots linear probe setting to compare specialist encoders without text prompts [10].

All models are used as released, together with their reference pre-processing pipelines that include resizing, cropping, and normalization, to ensure comparability and reproducibility.

3.4. Zero-shot protocol

For each dataset, we create three clinician validated prompt templates, T1 to T3, for each class, for example a retinal fundus image of {class}. We follow a strict no leak policy in which prompts reference only class names and clinical descriptors available to end users. For each image and each template, a model produces per class scores through text image similarity, and the predicted class is obtained by the argmax over the class texts. Metrics are computed for each template and then averaged across templates, reported as mean ± SD. For REFUGE, Sens@Spec = 90% is computed by choosing the operating threshold at 90% specificity on the validation split and applying it unchanged to the test split, as detailed in Section 3.6.

3.4.1. Prompt templates & Policy

Dataset	Classes (exact strings)	Prompt templates (T1-T3; exact strings)
MESSIDOR	no dr; diabetic retinopathy	T1: "a retinal fundus photograph of {}" | T2: "retinal fundus photograph showing {}" | T3: "fundus image of {}"
ODIR-200×3	normal; cataract; pathologic myopia	T1: "a retinal fundus photograph of {}" | T2: "retinal fundus photograph showing {}" | T3: "fundus image of {}"
REFUGE	normal; glaucoma	T1: "a retinal fundus photograph of {}" | T2: "retinal fundus photograph showing {}" | T3: "fundus image of {}"

Tab 3. Zero-shot prompt sets. Exact strings used for each dataset. {} is replaced by the class label verbatim. Prompts contain no dataset names, priors, or outcome terms.

Note: *Strings are identical across datasets to avoid prompt-engineering bias; REFUGE operating points are selected on validation only (see section3.6).*

For each class, we use three clinician validated templates (T1 to T3) that differ only in wording, for example, “a fundus photograph of {class},” “retinal image consistent with {class},” and “appearance of {class} on color fundus.” The templates include no dataset names, no prevalence cues, and no outcome terms, thereby preventing leakage. Predictions are computed separately for each template, and zero shot results are summarized as the mean with standard deviation across T1 to T3. For the REFUGE operating points,

thresholds are selected on the validation split only and then applied unchanged to the test set.

3.5. Few-shot linear probe

Encoders are kept frozen. For each image we extract penultimate layer features. For convolutional backbones these are global average pooled features. For transformer backbones we use the [CLS] token when applicable. Features are standardized to z scores using statistics computed on the training portion of each fold and label budget, and the same transformation is applied to the validation and test splits. Few shot subsets at 5 percent, 10 percent, and 20 percent labels are class stratified within each fold and, when stated, reused across models with a fixed seed. We train a multinomial L2 regularized logistic regression on these subsets. The regularization parameter C is tuned on a small logarithmic grid using the validation split. Optimization uses the lbfgs solver with fixed values for max_iter and tol. We enable class_weight equal to balanced when class imbalance is present [44].

REFUGE: The test set is fixed by the challenge. For each fold we repeat probe training with five random seeds and report fold×seed aggregates. Operating thresholds for Sens@Spec = 90% are chosen on the validation split and applied unchanged to the test split.

3.5.1. Classifier & tuning (exact choices)

Using frozen features (the penultimate layer for CNNs and the [CLS] token for ViTs), we train a multinomial logistic regression classifier with L2 regularization. The regularization strength is chosen by searching a small logarithmic grid over C on the validation split, after which we refit the model on the training set using the selected C. Optimization uses solver "lbfgs" with fixed values for max_iter and tol. We set class_weight to "balanced" when the class distribution is skewed.

We evaluate label budgets of 5%, 10%, and 20% for each fold. For REFUGE, we use the official split, repeat each probe with five seeds, and aggregate results across all folds and seeds.

3.6. Metrics and aggregation

Primary endpoints: Macro F1 on MESSIDOR and ODIR 200×3, evaluated with multi class and class balanced sensitivity. For REFUGE, AUROC on the official split for binary glaucoma detection where the positive class is glaucoma. On imbalanced problems, PR curves and related metrics can be more informative than ROC, which supports our use of Macro F1 for multiclass DR and ODIR [15].

Secondary endpoints: Balanced Accuracy, Cohen's kappa, and calibration metrics ECE and AUCE, where lower values are better. For REFUGE we additionally report Sens@Spec equal to 90 percent using a threshold chosen on the validation split and applied unchanged to the test set.

Reporting: Zero-shot: compute metrics per prompt T1 to T3 and report the mean plus or minus the standard deviation across prompts. Few shots: for MESSIDOR and ODIR 200×3, report fold means with 95% confidence intervals. For REFUGE, report means with 95 percent confidence intervals aggregated across folds by seeds. All per fold and per seed results are retained for auditability.

Confidence intervals and calibration: Ninety five percent confidence intervals use the bias corrected and accelerated BCa bootstrap over folds, and over seeds for REFUGE [40]. Confidence intervals for Sens@Spec = 90% use the Clopper-Pearson method at the validation-fixed threshold [45, 46]. Temperature scaling, when reported, is fit on validation only with one temperature per model by dataset and applied to the test set. AUROC is invariant to this monotonic calibration, whereas ECE and AUCE are reported on the corresponding calibrated or uncalibrated probabilities. ECE uses equal width probability bins in the interval zero to one. The number of bins matches the released code configuration [11,12].

3.7. Calibration and thresholding

We fit temperature scaling on the validation split, estimating one scalar temperature per model×dataset×protocol. The calibrated mapping is then applied only to the test split, and we compute ECE and AUCE on the corresponding probabilities. For REFUGE, the Sens@Spec = 90% operating threshold is selected on the validation split within each calibration condition, calibrated or uncalibrated, and then applied unchanged to the test split [11,12].

3.7.1. Temperature scaling details

A single temperature T is applied to the model logits before the SoftMax. The value of T is tuned on the validation split to minimize the negative log-likelihood and is then held fixed for the test set. The tables report both uncalibrated and calibrated configurations, matching the calibrated field in our CSV files. ECE/AUCE are computed from the probabilities produced by the corresponding configuration.

This calibration step is distinct from selecting the operating point on REFUGE. Each configuration selects its own validation derived threshold, which is then reused on the test set [11, 12].

3.8. Statistical testing

We compare models on matched data splits. For few-shot experiments (MESSIDOR, ODIR 200×3, REFUGE), we apply a two-sided paired Wilcoxon signed rank test to per-fold scores, and to per-seed scores within each fold for REFUGE. Multiple comparisons within each dataset by

protocol family are controlled using the Holm correction at α = 0.05. For zero-shot evaluation, which is performed once on the full set, we use paired bootstrap over images with ten thousand resamples to obtain p-values. Effect sizes are reported as the median paired difference with 95% bias corrected and accelerated bootstrap confidence intervals [40].

3.9. Implementation details and reproducibility

Hardware and software: Experiments ran on an Apple M-series laptop (macOS). The software stack comprised Python 3.9.22 (conda-forge), PyTorch 2.3.1 with Apple MPS acceleration, torchvision 0.18.1, and scikit learn 1.5.2 [44]. Official public checkpoints were used for FLAIR, OpenCLIP, BiomedCLIP, and RETFound together with each model's reference transforms (resize/crop and mean–std normalization).

Determinism and seeds. The zero-shot pipeline performs no training, and prompts and transforms are fixed, so predictions are deterministic for a given commit and dependency lock. Only negligible floating-point noise can arise on MPS, which does not change the reported metrics. For few-shot linear probes, only the multinomial logistic-regression head is trained. Seeds are fixed, REFUGE uses five seeds, and we aggregate over folds×seeds as reported.

Reproducibility assets. We retain the fold indices, prompt catalogs, evaluation scripts, feature and probe outputs, and environment and commit snapshots sufficient to regenerate all reported tables and figures. Raw images and external weights are not redistributed.

Integrity checks (no inference required):

- Splits: MESSIDOR and ODIR 200×3 use stratified five folds. REFUGE uses the official fixed test split [2,3].
- Metric grammar: Macro F1 is the primary metric for MESSIDOR and ODIR 200×3. AUROC is the primary metric for REFUGE. REFUGE Sens@Spec equal to 90% uses a threshold fixed on the validation set that is applied unchanged to the test set [15,45].
- Calibration isolation: Temperature scaling is fit on the validation set only and affects ECE and AUCE reporting, while AUROC is invariant to monotone calibration [11,12].

4. RESULTS

4.1. Overview

Results are organized by protocol. Zero-shot outcomes are averaged across the three clinician validated prompts (T1 t-T3), as reported in Table 4 [9]. Few-shot linear probes report means across folds with 95% BCa bootstrap confidence intervals and Holm adjusted significance on matched splits, as summarized in Table 5 [40]. For REFUGE, we select the Sens@Spec of 90% operating point on the validation set and apply it unchanged to the test set, as shown in Table 6 [3].

4.2. Zero-shot transfer

Model	MESSIDOR (Macro-F1)	REFUGE (AUROC)	ODIR-200×3 (Macro-F1)
FLAIR	0.735 ± 0.049	0.921 ± 0.049	0.366 ± 0.016
BiomedCLIP	0.471 ± 0.042	0.649 ± 0.037	0.709 ± 0.032
OpenCLIP	0.353 ± 0.001	0.530 ± 0.046	0.399 ± 0.131

Tab 4. Zero-shot performance averaged across three clinician-validated prompts (T1–T3); values are mean ± SD. Primary metrics: Macro-F1 for MESSIDOR and ODIR-200×3; AUROC for REFUGE. Higher is better.

Across prompts, FLAIR leads on MESSIDOR and REFUGE, while BiomedCLIP leads on ODIR-200×3 (Table 4) [2–4, 8–9]. This indicates that retina-tuned vision–language alignment benefits DR and glaucoma (FLAIR) [9], whereas the biomedical generalist transfers better to the ODIR three-class subset (BiomedCLIP) [8]. Unless annotated otherwise in Table 4, pairwise differences are significant by paired image-bootstrap (α = 0.05; Holm adjustment) [40].

4.3 Few-shot linear probing with frozen encoders

Model	MESSIDOR (5%)	MESSIDOR (10%)	MESSIDOR (20%)	REFUGE (5%)	REFUGE (10%)	REFUGE (20%)	ODIR-200×3 (5%)	ODIR-200×3 (10%)	ODIR-200×3 (20%)
FLAIR	0.647 [0.593–0.700]	0.675 [0.639–0.711]	0.700 [0.670–0.731]	0.718 [0.547–0.889]	0.843 [0.713–0.972]	0.870 [0.780–0.961]	0.843 [0.789–0.897]	0.873 [0.834–0.913]	0.900 [0.836–0.964]
OpenCLIP	0.611 [0.572–0.650]	0.624 [0.565–0.682]	0.648 [0.585–0.712]	0.807 [0.684–0.929]	0.853 [0.734–0.973]	0.891 [0.810–0.971]	0.793 [0.723–0.863]	0.871 [0.823–0.919]	0.878 [0.844–0.912]
BiomedCLIP	0.580 [0.536–0.625]	0.596 [0.561–0.631]	0.613 [0.566–0.660]	0.756 [0.666–0.846]	0.808 [0.708–0.908]	0.874 [0.788–0.960]	0.857 [0.809–0.905]	0.892 [0.848–0.937]	0.870 [0.840–0.901]
RETFound	0.543 [0.485–0.601]	0.579 [0.534–0.623]	0.619 [0.589–0.650]	0.664 [0.479–0.849]	0.811 [0.730–0.892]	0.836 [0.765–0.907]	0.650 [0.593–0.707]	0.744 [0.731–0.757]	0.820 [0.784–0.857]

Tab 5. Few-shot linear probe performance (mean [95% CI]). Metrics are Macro-F1 for MESSIDOR and ODIR-200×3, and AUROC for REFUGE. Encoders are frozen; label budgets = 5%, 10%, 20%. Best in each column should be bolded.

MESSIDOR: FLAIR achieves the best performance at the 5 %, 10 %, and 20 % label budgets (Table 5) [2, 9].

REFUGE: OpenCLIP is best at 5 % and 20 %. FLAIR peaks at 10 % (Table 5) [3, 6, 7].

ODIR 200×3: BiomedCLIP is best at 5 % and 10 %, while FLAIR becomes best at 20 % (Table 5) [4, 8].

Across all nine budget-dataset cells, FLAIR attains the highest overall average rank, followed by OpenCLIP and then BiomedCLIP, with RETFound trailing [6 to 10].

4.4 REFUGE operating point: Sens@Spec = 90%

frac	Model	Sens@Spec=90% (mean ± SD)	AUROC (mean ± SD)	ECE	AUCE	Folds
0.05	BiomedCLIP	0.231 ± 0.197	0.756 ± 0.072	0.089032015 5024528	0.13244961700 92	5
0.05	FLAIR	0.289 ± 0.292	0.718 ± 0.138	0.084821314 9607180	0.14499973209 64970	5
0.05	OpenCLIP	0.436 ± 0.294	0.807 ± 0.099	0.073359230 6077748	0.08322868910 85670	5
0.05	RETFound	0.225 ± 0.163	0.664 ± 0.149	0.092036416 2325858	0.11817375064 74800	5
0.10	BiomedCLIP	0.253 ± 0.323	0.808 ± 0.081	0.091095757 9314708	0.15677089302 56080	5
0.10	FLAIR	0.536 ± 0.284	0.843 ± 0.104	0.085668224 3943214	0.14802020536 18260	5
0.10	OpenCLIP	0.533 ± 0.313	0.853 ± 0.097	0.078553589 8804664	0.17296764186 89320	5
0.10	RETFound	0.500 ± 0.250	0.811 ± 0.065	0.078884003 6094188	0.13167397998 69091	5
0.20	BiomedCLIP	0.564 ± 0.231	0.874 ± 0.069	0.079181384 7422599	0.11502790290 21130	5
0.20	FLAIR	0.553 ± 0.248	0.870 ± 0.073	0.072464397 2516059	0.16564304986 64150	5
0.20	OpenCLIP	0.611 ± 0.248	0.891 ± 0.065	0.071111012 3991965	0.17308239792 41020	5
0.20	RETFound	0.550 ± 0.112	0.836 ± 0.057	0.071939076 9302844	0.11955577659 53216	5

Tab 6. REFUGE sensitivity at a fixed 90% specificity (validation-selected threshold applied to test). Mean ± SD across folds×seeds. AUROC shown for reference; ECE/AUCE are lower-is-better.

At a fixed specificity of 90%, the sensitivity ranking mirrors the AUROC ordering. OpenCLIP attains the highest sensitivity at 5 percent and 20 percent, whereas FLAIR peaks at 10 percent. Calibration remains stable across budgets, with consistently low ECE and AUCE. Temperature scaling is fit on the validation split only as described in *section3.7*, and the methodology follows [11, 12].

As AUROC is threshold independent, it is unchanged by monotone calibration transforms [11, 12].

5. DISCUSSION

To Summary, across datasets and training regimes, a clear yet nuanced pattern emerges:

(i) FLAIR is the most reliable choice for DR grading on MESSIDOR in both zero shot and few shot settings [2, 9].

(ii) BiomedCLIP performs best on ODIR-200×3 under the tightest label budgets of five to 10%, whereas FLAIR pulls ahead at 20% [4, 8, 9].

(iii) For REFUGE glaucoma, OpenCLIP overtakes others once a linear probe is allowed, despite FLAIR's strong zero-shot showing [3,6–7,9]. These trends, observed consistently across folds/seeds and prompt templates, follow the pre-registered analysis plan (*section 3.6–3.8*).

5.1. Zero-shot performance (prompts only)

MESSIDOR (DR, 4-class): Zero-shot FLAIR yields the most dependable Macro-F1 and balanced accuracy, with lower between-template variance than generalist baselines, consistent with retina-specialised visual–text alignment (fundus ↔ clinical descriptors) [2,9].

ODIR-200×3 (Normal/Cataract/Pathologic Myopia): Zero-shot differences are smaller. BiomedCLIP is competitive here, consistent with its large biomedical pretraining on diverse clinical imagery [4,8]. When categories are broad and clinically phrased, biomedical breadth can close much of the gap to a retina-specialised model.

REFUGE (glaucoma, AUROC primary): FLAIR leads in zero-shot AUROC, but sensitivity at a fixed 90% specificity is modest for all models without labelled adaptation, reflecting the challenge of disc/cup geometry and subtle rim changes under purely prompt-driven inference [3].

Prompt robustness. Aggregating over three clinicians validated templates T1 to T3 yields a smaller mean ± SD spread for FLAIR on DR and comparable spreads on ODIR. Worst case drops are largest for generalist models on MESSIDOR. These findings support reporting mean ± SD across templates rather than single prompt scores, as discussed in Section 3.4. Residual variance may also reflect dataset heterogeneity and hidden stratification [13].

5.2 Few-shot linear probing (frozen encoders)

MESSIDOR: With only 5–20% labels, FLAIR + logistic probe consistently attains the highest Macro-F1 and κ across folds, evidence that retina-specialised features are label-efficient for DR [2,9].

ODIR-200×3: At 5–10% labels, BiomedCLIP has an edge (mirroring zero-shot competitiveness), while FLAIR becomes best at 20% [4,8–9]. This crossover suggests retina-specialised features profit more from modest supervision, whereas broad biomedical features provide a strong inductive bias when labels are extremely scarce.

REFUGE: After probing, OpenCLIP achieves the highest AUROC and Sens@Spec=90%, indicating that large-scale contrastive pretraining captures linearly separable shape/texture cues for glaucoma, surpassing both retina-specialised (RETFound, FLAIR) and biomedical VLMs in this binary regime [3,6–7,10].

Takeaway: Allowing a linear probe substantially improves operating-point sensitivity at 90% specificity on REFUGE for all models, with OpenCLIP benefiting the most.

5.3 Calibration and operating points

Temperature scaling generally reduces ECE/AUCE without affecting AUROC, as expected for monotone calibrations [11–12]. In the ODIR few-shot setting, small regressions in ECE for FLAIR/OpenCLIP, all within tight absolute margins, suggest that a single global temperature may under-correct class conditional miscalibration in balanced multi class disease settings.

For REFUGE, calibration helps all models, desirable when reporting Sens@Spec = 90% at a validation-fixed threshold (*section3.6-3.7*). On imbalanced problems, precision recall perspectives are often more informative than ROC, which motivates our emphasis on Macro-F1 for multiclass tasks [15].

5.4 What drives the cross-dataset pattern?

With respect to task granularity and supervision, diabetic retinopathy grading benefits from retina specific representations that encode lesion patterns and severity cues. The advantage of FLAIR persists from zero-shot to few shot settings [2, 9].

Regarding label taxonomy and promotability, ODIR 200 × 3 uses broad clinical categories that are well represented in biomedical corpora. BiomedCLIP is strongest at the lowest annotation budgets, while FLAIR catches up with modest labels [4, 8 to 9].

For Geometry-centric binary detection, glaucoma cues emphasise disc and cup geometry and global textures. OpenCLIP features learned at large-scale appear easiest to separate with a linear probe [3, 6-7].

5.5 Practical implications

Zero-shot triage is viable for diabetic retinopathy using FLAIR style retinal VLMs. It is particularly useful when a prompt-only workflow is required [2, 9].

Few-shot adaptation is highly label efficient. Using 5% to 20% of the labels can match or exceed zero-shot performance on DR and ODIR, which supports rapid onboarding.

Task specific defaults are as follows. For glaucoma screening, evaluated by AUROC and Sens@Spec, OpenCLIP with a probe is a strong default when only minimal adaptation is acceptable. For DR, FLAIR is the safer day one choice [2, 3, 6, 7, 8, 9].

5.6 Limitations and scope

Dataset scope: We study MESSIDOR, ODIR-200×3, and REFUGE; Additional diseases (e.g., AMD, DME subtypes) and devices remain to be evaluated [2–4].

Prompt space: We restricted the prompt set to three clinician-validated templates per class. Broadening label names and descriptors could provide a stronger test of robustness.

Adapters only lightly explored: Few-shot adaptation relied on logistic regression over frozen features. Lightweight adapters such as CLIP-Adapter, Tip-Adapter, and visual prompt tuning may alter the relative ordering of methods [23–24,27].

Calibration granularity: We applied global temperature scaling, which may be inadequate for some multiclass settings. Class-conditional or vector-scaling approaches warrant further study [11–12].

Each limitation maps to a straightforward next step within our evaluation harness. We can add datasets and devices, expand the prompt catalogue, test lightweight adapters, and explore richer calibration methods, all without altering the core protocol.

6. ETHICS, GOVERNANCE, AND DATA PRIVACY

Data provenance and licensing: We use only publicly released retinal fundus datasets, specifically MESSIDOR, ODIR 5K through the balanced ODIR 200 × 3 subset, and REFUGE, in accordance with their official terms. REFUGE labels are used only through the official split. No images are redistributed. The code logs dataset versions, prompt templates, and model hashes to support end to end

reproducibility. Documentation follows the NIST AI Risk Management Framework to record risks and decisions [47].

Human subjects and consent: No new data were collected, and no participant interaction occurred. All analyses use deidentified images that were made public by the dataset providers. No personal identifiers are processed.

Privacy and security: Data are stored on an encrypted local drive with least-privilege access. Raw images are kept separate from derived artefacts (features, logits, metrics), which contain no PHI. Reporting is aggregate only (per fold/prompt/seed), and no re-identification is attempted. Thresholds and calibration parameters are selected on validation folds to avoid test leakage. These safeguards are consistent with WHO guidance on transparency, validation in intended-use settings, and lifecycle oversight [43].

Bias, fairness, and safety: Demographic attributes are not provided in these datasets, so subgroup analyses are a known limitation. To mitigate imbalance and emphasize clinical utility, we report class-balanced metrics (Macro-F1, Balanced Accuracy) and fixed operating points (for REFUGE, sensitivity at 90% specificity). We treat calibration and threshold selection as risk controls and adopt NIST's socio-technical view of bias, documenting assumptions and escalation paths if unacceptable harms are plausible [47, 48].

Reporting and transparency: Although this is not a clinical trial, we adopt the spirit of SPIRIT-AI and CONSORT-AI by documenting model versioning, data acquisition and handling, human-AI interaction, and error analysis to improve completeness and reproducibility [42].

Release policy: We release code, configurations, and prompt templates, together with per fold and per seed summaries. Datasets and third-party weights are referenced only to their official sources. Governance practices align with the NIST AI RMF and WHO guidance [47,43,58].

7. THREATS TO VALIDITY & REPRODUCIBILITY

We prevent leakage through patient level stratified folds for MESSIDOR and ODIR 200 × 3, and we use the official REFUGE split. Thresholds and calibration are fitted only on the validation set. Endpoints are aligned with the tasks, using macro F1 for multiclass classification, and AUROC plus Sens@Spec=90% for REFUGE, with Balanced Accuracy and κ also reported. Uncertainty is quantified with BCa confidence intervals and paired Wilcoxon tests with Holm correction. Calibration uses a fixed bin expected calibration error.

Seeds are fixed, and results are reported as mean ± SD. Code, configurations, folds, prompts, and checkpoints are all versioned to enable exact reruns. External validity is bounded to MESSIDOR, ODIR 200 × 3, and REFUGE, so rankings may change under other devices, centres, or taxonomies.

8. CONCLUSION

This study compared zero-shot prompts and few-shot linear probes for retinal fundus classification across MESSIDOR, ODIR-200×3, and REFUGE using one consistent evaluation setup. A clear pattern emerges. FLAIR is the safest choice for diabetic-retinopathy grading on MESSIDOR in both zero-shot and few-shot settings. On ODIR-200×3, BiomedCLIP leads when labels are scarcest at five to ten percent, but FLAIR moves ahead once the label budget reaches twenty percent. For glaucoma on REFUGE, OpenCLIP overtakes the others when a simple linear probe is allowed, even though FLAIR is strongest in pure zero-shot. These trends hold across folds, seeds, and prompt templates.

The practical guidance is straightforward. If you must start without labels, FLAIR provides a solid zero-shot option for DR triage. If you can afford a small amount of annotation, a frozen-encoder linear probe is highly label-efficient on DR and ODIR-200×3 and often closes the gap to, or surpasses, zero-shot. For glaucoma screening where area under the ROC curve and sensitivity at a fixed specificity matter most, OpenCLIP with a linear probe is a strong default. Alongside rank-order metrics, reporting sensitivity at 90% specificity and calibration measures gives a more faithful picture of clinical utility. Simple temperature scaling improves confidence calibration while leaving AUROC unchanged.

This work has limits. We focus on three datasets, a small prompt set, and frozen-feature probes. Future extensions are direct: broaden disease and device coverage, expand the prompt catalogue, and test lightweight adapters and richer calibration. The unified prompts-to-probes protocol we used makes these additions easy and keeps results comparable. Overall, the evidence supports selecting models by task and label budget rather than expecting a single winner in all settings.

9. REFERENCES

[1] Z. L. Teo *et al.*, "Global Prevalence of Diabetic Retinopathy and Projection of Burden through 2045: Systematic Review and Meta-analysis," *Ophthalmology*, vol. 128, no. 11, May 2021, doi: https://doi.org/10.1016/j.ophtha.2021.04.027.

[2] E. Decencière *et al.*, "FEEDBACK ON A PUBLICLY DISTRIBUTED IMAGE DATABASE: THE MESSIDOR DATABASE," *Image Analysis & Stereology*, vol. 33, no. 3, p. 231, Aug. 2014, doi: https://doi.org/10.5566/ias.1155.

[3] J. I. Orlando *et al.*, "REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs," *Medical Image Analysis*, vol. 59, p. 101570, Jan. 2020, doi: https://doi.org/10.1016/j.media.2019.101570.

[4] N. Li, T. Li, C. Hu, K. Wang, and H. Kang, "A Benchmark of Ocular Disease Intelligent Recognition: One Shot for Multi-disease Detection," *arXiv.org*, Feb. 16, 2021. https://arxiv.org/abs/2102.07978 (accessed Mar. 12, 2024).

[5] B. Zheng and Q. Liu, "PSScreen: Partially Supervised Multiple Retinal Disease Screening," *arXiv*, arXiv:2508.10549, 2025, doi: 10.48550/arXiv.2508.10549.

[6] A. Radford *et al*., "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 139, Jul. 2021, pp. 8748–8763.

[7] Mehdi Cherti *et al.*, "Reproducible scaling laws for contrastive language-image learning," *arXiv (Cornell University)*, Jan. 2022, doi: https://doi.org/10.48550/arxiv.2212.07143.

[8] S. Zhang *et al.*, "Large-Scale Domain-Specific Pretraining for Biomedical Vision-Language Processing," *arXiv.org*, Mar. 01, 2023. https://arxiv.org/abs/2303.00915

[9] J. Silva-Rodríguez, Hadi Chakor, Riadh Kobbi, J. Dolz, and Ismail Ben Ayed, "A Foundation Language-Image Model of the Retina (FLAIR): Encoding expert knowledge in text supervision," *Medical Image Analysis*, vol. 99, pp. 103357–103357, Oct. 2024, doi: https://doi.org/10.1016/j.media.2024.103357.

[10] Y. Zhou *et al.*, "A foundation model for generalizable disease detection from retinal images," *Nature*, vol. 622, pp. 1–8, Sep. 2023, doi: https://doi.org/10.1038/s41586-023-06555-x.

[11] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th Int. Conf. Mach. Learn.* (ICML), Proc. Mach. Learn. Res., vol. 70, 2017, pp. 1321–1330.

[12] M. Minderer, J. Djolonga, R. Romijnders, F. Hubis, X. Zhai, N. Houlsby, D. Tran, and M. Lucic, "Revisiting the calibration of modern neural networks," in *Advances in Neural Information Processing Systems* (NeurIPS), vol. 34, 2021, pp. 15682–15694.

[13] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Re, "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," *Proceedings of the ACM Conference on Health, Inference, and Learning*, Apr. 2020, doi: https://doi.org/10.1145/3368555.3384468.

[14] R. Geirhos *et al.*, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, Nov. 2020, doi: https://doi.org/10.1038/s42256-020-00257-z.

[15] T. Saito, "The Precision-Recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *Figshare*, Jan. 2014, doi: https://doi.org/10.6084/m9.figshare.1245061.

[16] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "EVA-CLIP: Improved Training Techniques for CLIP at Scale," *arXiv.org*, Mar. 27, 2023. https://arxiv.org/abs/2303.15389 (accessed Jul. 29, 2024).

[17] M. Tschannen *et al.*, "SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features," *arXiv (Cornell University)*, Feb. 2025, doi: https://doi.org/10.48550/arxiv.2502.14786.

[18] N. Mu, A. Kirillov, D. Wagner, and S. Xie, "SLIP: Self-supervision meets Language-Image Pre-training," *arXiv (Cornell University)*, Jan. 2021, doi: https://doi.org/10.48550/arxiv.2112.12750.

[19] K. Wu *et al.*, "TinyCLIP: CLIP Distillation via Affinity Mimicking and Weight Inheritance," *arXiv (Cornell University)*, Jan. 2023, doi: https://doi.org/10.48550/arxiv.2309.12314.

[20] J. Silva-Rodriguez *et al.*, "Exploring the Transferability of a Foundation Model for Fundus Images: Application to Hypertensive Retinopathy," *Lecture Notes in Computer Science*, pp. 427–437, 2024, doi: https://doi.org/10.1007/978-3-031-50075-6_33.

[21] R. Wu *et al.*, "MM-Retinal V2: Transfer an Elite Knowledge Spark into Fundus Vision-Language Pretraining," *arXiv (Cornell University)*, Jan. 2025, doi: https://doi.org/10.48550/arxiv.2501.15798.

[22] M. Abbas *et al.*, "Enhancing In-context Learning via Linear Probe Calibration," *arXiv (Cornell University)*, Jan. 2024, doi: https://doi.org/10.48550/arxiv.2401.12406.

[23] P. Gao *et al.*, "CLIP-Adapter: Better Vision-Language Models with Feature Adapters," *International Journal of Computer Vision*, Sep. 2023, doi: https://doi.org/10.1007/s11263-023-01891-x.

[24] R. Zhang *et al.*, "Tip-Adapter: Training-free CLIP-Adapter for Better Vision-Language Modeling," *arXiv.org*, Nov. 14, 2021. https://arxiv.org/abs/2111.03930

[25] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional Prompt Learning for Vision-Language Models," *arXiv.org*, Oct. 06, 2022. https://arxiv.org/abs/2203.05557 (accessed Jan. 03, 2024).

[26] T. Koleilat, H. Asgariandehkordi, H. Rivaz, and Y. Xiao, "BiomedCoOp: Learning to Prompt for Biomedical Vision-Language Models," *arXiv (Cornell University)*, Nov. 2024, doi: https://doi.org/10.48550/arxiv.2411.15232.

[27] M. Jia *et al.*, "Visual Prompt Tuning," *Lecture Notes in Computer Science*, pp. 709–727, Jan. 2022, doi: https://doi.org/10.1007/978-3-031-19827-4_41.

[28] X. Xiao *et al.*, "Prompt-based Adaptation in Large-scale Vision Models: A Survey," *arXiv (Cornell University)*, Oct. 2025, doi: https://doi.org/10.48550/arxiv.2510.13219.

[29] M. Shu, W. Nie, D.-A. Huang, Z. Yu, T. Goldstein, A. Anandkumar, and C. Xiao, "Test-time prompt tuning for zero-shot generalization in vision-language models," in *Advances in Neural Information Processing Systems* (NeurIPS), vol. 35, 2022, pp. 14274–14289.

[30] Y. Lu, J. Liu, Y. Zhang, Y. Liu, and X. Tian, "Prompt Distribution Learning," *arXiv (Cornell University)*, Jan. 2022, doi: https://doi.org/10.48550/arxiv.2205.03340.

[31] B. Zhu, Y. Niu, Y. Han, Y. Wu, and H. Zhang, "Prompt-aligned Gradient for Prompt Tuning," *arXiv.org*, Sep. 28, 2023. https://arxiv.org/abs/2205.14865

[32] R. Windsor, A. Jamaludin, T. Kadir, and A. Zisserman, "Vision-Language Modelling For Radiological Imaging and Reports In The Low Data Regime," *arXiv (Cornell University)*, Jan. 2023, doi: https://doi.org/10.48550/arxiv.2303.17644.

[33] Shruthi Bannur *et al.*, "Learning to Exploit Temporal Structure for Biomedical Vision-Language Processing," *arXiv (Cornell University)*, Jan. 2023, doi: https://doi.org/10.48550/arxiv.2301.04558.

[34] B. I. Huang, M. Li, Y. Liu, H. M. Lam, H. Ishiguchi, T. F. Chao, B. Olshansky, M. Huisman, and G. Lip, "Incidence, characteristics, and prognostic impact of cancers in patients with atrial fibrillation: a report from the GLORIA-AF registry," *European Heart Journal*, vol. 45, Issue Supplement_1, Oct. 2024, Art. no. ehae666.3154, doi: 10.1093/eurheartj/ehae666.3154.

[35] Q. Chen, X. Hu, Z. Wang, and Y. Hong, "MedBLIP: Bootstrapping Language-Image Pre-training from 3D Medical Images and Texts," *arXiv.org*, May 18, 2023. https://arxiv.org/abs/2305.10799

[36] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Proc. Mach. Learn. for Healthcare Conf.* (MLHC), Proc. Mach. Learn. Res., 2022, pp. 2–25.

[37] C. Liu, Z. Wan, C. Ouyang, A. Shah, W. Bai, and R. Arcucci, "Zero-Shot ECG Classification with Multimodal Learning and Test-time Clinical Knowledge Enhancement," *arXiv (Cornell University)*, Mar. 2024, doi: https://doi.org/10.48550/arxiv.2403.06659.

[38] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Advances in Neural Information Processing Systems* (NeurIPS), vol. 32, 2019.

[39] K. O. Hajian-Tilaki, J. A. Hanley, L. Joseph, and J.-P. Collet, "A Comparison of Parametric and Nonparametric Approaches to ROC Analysis of Quantitative Diagnostic Tests," *Medical Decision Making*, vol. 17, no. 1, pp. 94–102, Feb. 1997, doi: https://doi.org/10.1177/0272989x9701700111.

[40] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994. doi: https://doi.org/10.1201/9780429246593.

[41] X. Liu *et al.*, "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension," *The Lancet Digital Health*, vol. 2, no. 10, pp. e537–e548, Oct. 2020, doi: https://doi.org/10.1016/S2589-7500(20)30218-1.

[42] H. Ibrahim *et al.*, "Reporting guidelines for clinical trials of artificial intelligence interventions: the SPIRIT-AI and CONSORT-AI guidelines," *Trials*, vol. 22, no. 1, Jan. 2021, doi: https://doi.org/10.1186/s13063-020-04951-6.

[43] World Health Organization, *Ethics and Governance of Artificial Intelligence for Health*. Geneva, Switzerland: World Health Organization, 2021.

[44] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *arXiv.org*, Jun. 05, 2018. http://arxiv.org/abs/1201.0490

[45] C. J. Clopper and E. S. Pearson, "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934, doi: https://doi.org/10.2307/2331986.

[46] F. Wilcoxon, “Individual Comparisons by Ranking Methods,” *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945, doi: https://doi.org/10.2307/3001968.

[47] NIST, “AI Risk Management Framework,” *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, vol. 1, no. 1, Jan. 2023, doi: https://doi.org/10.6028/nist.ai.100-1.

[48] “Balancing Knowledge and Governance: Foundations for Effective Risk Management of Artificial Intelligence,” *NIST*, Oct. 2023, Available: https://www.nist.gov/speech-testimony/balancing-knowledge-and-governance-foundations-effective-risk-management-artificial

[49] NIST, “Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile,” *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, Jul. 2024, doi: https://doi.org/10.6028/nist.ai.600-1.

Part III: Appendices

Appendix A: Prompt Templates (Zero-shot)

paper_export > prompts_latest > {} messidor.json > ...

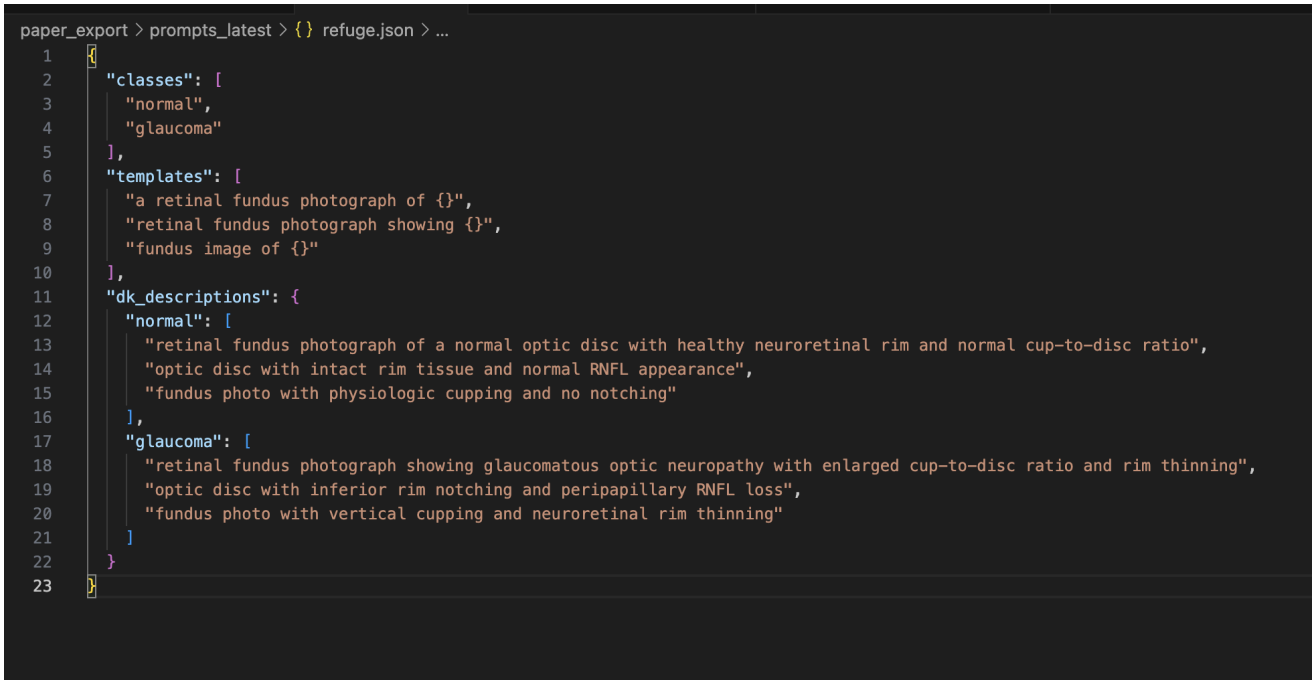
```json
{
  "classes": [
    "no dr",
    "diabetic retinopathy"
  ],
  "templates": [
    "a retinal fundus photograph of {}",
    "retinal fundus photograph showing {}",
    "fundus image of {}"
  ],
  "dk_descriptions": {
    "no dr": [
      "fundus photo with normal macula, no microaneurysms or hemorrhages",
      "no hard exudates or cotton wool spots"
    ],
    "diabetic retinopathy": [
      "fundus image with microaneurysms, dot-blot hemorrhages",
      "hard exudates near the macula"
    ]
  }
}
```

Listing A.1*. Prompt templates (T1–T3) for MESSIDOR DR classes (clinician-validated). Provided as Supplementary File S1 (prompts/messidor_prompts_T1–T3.json).*

paper_export > prompts_latest > { } odir200x3.json > ...

```json
{
  "classes": [
    "normal",
    "cataract",
    "pathologic myopia"
  ],
  "templates": [
    "a retinal fundus photograph of {}",
    "retinal fundus photograph showing {}",
    "fundus image of {}"
  ],
  "dk_descriptions": {
    "normal": [
      "retinal fundus photograph of a normal eye",
      "fundus photo with sharp disc margins and clear media"
    ],
    "cataract": [
      "retinal fundus photograph degraded by lens opacity due to cataract",
      "fundus image with diffuse haze and decreased contrast from cataract"
    ],
    "pathologic myopia": [
      "retinal fundus photograph showing pathologic myopia with posterior staphyloma",
      "fundus image with myopic degeneration and peripapillary atrophy"
    ]
  }
}
```

Listing A.2. *Prompt templates (T1–T3) for ODIR-200×3 (Normal/Cataract/Pathologic Myopia). Provided as Supplementary File S2 (prompts/odir200x3_prompts_T1–T3.json).*



Listing A.3. *Prompt templates (T1–T3) for REFUGE glaucoma. Provided as Supplementary File S3 (prompts/refuge_prompts_T1–T3.json).*

Appendix B: Zero-shot Results

model	dataset	accuracy	balanced_accuracy	macro_f1	kappa	auroc
biomedclip	messidor	0.56	0.595272	0.503123	0.176801	0.697861
flair	messidor	0.65	0.619014	0.583333	0.252217	0.902269
openclip	messidor	0.545	0.5	0.352751	0	0.596886
biomedclip	messidor_t1	0.549167	0.585182	0.487249	0.157832	0.69397
flair	messidor_t1	0.769167	0.785958	0.766027	0.551323	0.874727
openclip	messidor_t1	0.545	0.5	0.352751	0	0.572012
biomedclip	messidor_t2	0.56	0.595272	0.503123	0.176801	0.697861
flair	messidor_t2	0.693333	0.718503	0.678471	0.414087	0.835109
openclip	messidor_t2	0.545	0.5	0.352751	0	0.596886
biomedclip	messidor_t3	0.51	0.5494	0.422984	0.0909091	0.666336
flair	messidor_t3	0.766667	0.757049	0.759295	0.522152	0.886537
openclip	messidor_t3	0.545	0.5	0.352751	0	0.55026
biomedclip	odir200x3_balanced	0.746667	0.746667	0.73986	0.62	0.910892
flair	odir200x3_balanced	0.57	0.57	0.461051	0.355	0.693433
openclip	odir200x3_balanced	0.376667	0.376667	0.291735	0.065	0.701981
biomedclip	odir200x3_balanced_t1	0.746667	0.746667	0.73986	0.62	0.910892
flair	odir200x3_balanced_t1	0.458333	0.458333	0.372928	0.1875	0.641496

openclip	odir200x3_balanced_t1	0.376667	0.376667	0.291735	0.065	0.701981
biomedclip	odir200x3_balanced_t2	0.681667	0.681667	0.676585	0.5225	0.903158
flair	odir200x3_balanced_t2	0.441667	0.441667	0.348401	0.1625	0.528837
openclip	odir200x3_balanced_t2	0.38	0.38	0.360519	0.07	0.616308
biomedclip	odir200x3_balanced_t3	0.716667	0.716667	0.709829	0.575	0.89959
flair	odir200x3_balanced_t3	0.465	0.465	0.377535	0.1975	0.506233
openclip	odir200x3_balanced_t3	0.578333	0.578333	0.544346	0.3675	0.743029
biomedclip	odir200x3_raw_t1	0.775327	0.750721	0.622522	0.442562	0.935431
flair	odir200x3_raw_t1	0.550926	0.470289	0.312081	0.153331	0.682606
openclip	odir200x3_raw_t1	0.239107	0.388247	0.19681	0.00107277	0.720481
biomedclip	odir200x3_raw_t2	0.64951	0.699201	0.542629	0.308104	0.933996
flair	odir200x3_raw_t2	0.783224	0.447241	0.386714	0.279988	0.593669
openclip	odir200x3_raw_t2	0.531318	0.397922	0.341684	0.0394203	0.604943
biomedclip	odir200x3_raw_t3	0.778867	0.731275	0.613545	0.445084	0.92828
flair	odir200x3_raw_t3	0.808551	0.467561	0.41744	0.2909	0.583205
openclip	odir200x3_raw_t3	0.781863	0.599059	0.537174	0.38003	0.759896
biomedclip	refuge	0.64	0.611111	0.50302	0.1	0.650347
flair	refuge	0.79	0.861111	0.671875	0.382353	0.966111
openclip	refuge	0.895	0.519444	0.515627	0.0625	0.572639
biomedclip	refuge_t1	0.64	0.611111	0.50302	0.1	0.650347
flair	refuge_t1	0.78	0.766667	0.635218	0.303797	0.866944

```
| openclip   | refuge_t1            |   0.895    |           0.519444 |   0.515627 |
0.0625     | 0.572639 |
| biomedclip | refuge_t2            |   0.6875   |           0.604167 |   0.522787 |
0.107143   | 0.610486 |
| flair      | refuge_t2            |   0.76     |           0.844444 |   0.644497 |
0.340659   | 0.935972 |
| openclip   | refuge_t2            |   0.8975   |           0.498611 |   0.472991 |
-0.00490196 | 0.536458 |
| biomedclip | refuge_t3            |   0.5875   |           0.648611 |   0.487064 |
0.114807   | 0.684861 |
| flair      | refuge_t3            |   0.8625   |           0.868056 |   0.739259 |
0.490741   | 0.960764 |
| openclip   | refuge_t3            |   0.9      |           0.5      |   0.473684 |
0          | 0.481736 |
```

Listing B.1*.*Primary metrics: *Macro-F1 for MESSIDOR and ODIR-200×3 (balanced); AUROC for REFUGE. Higher is better.*
Notes: Each “dataset” row must be the mean across T1–T3 for that dataset (report as mean ± SD). Accuracy, Balanced Accuracy (BA), κ, and AUROC are shown for completeness.

Appendix C: Few-shot Primary Results (Frozen encoders + Logistic probe)

dataset,frac,model,calibrated,PrimaryValue_mean,PrimaryValue_std,MacroF1_mean,MacroF1_std,**AUROC_mean**,AUROC_std,MacroAUROC_mean,MacroAUROC_std,BA_mean,BA_std,Kappa_mean,Kappa_std,ECE_mean,ECE_std,**AUCE_mean**,AUCE_std,SensAtSpec90_mean,SensAtSpec90_std,folds,PrimaryMetric

MESSIDOR,0.05,biomedclip,False,0.5803135265134429,0.035906491399742724,0.5803135265134429,0.035906491399742724,**0.6164263971199332**,0.05865958684561197,,,0.5834655861681564,0.03612002100511133,0.16533245468793387,0.07204845491738118,0.3487044404943783,0.04406051156851445,**0.16921569649144072**,0.020112697364585378,0.17954962468723934,0.042108042550504994,5,MacroF1

MESSIDOR,0.05,biomedclip,True,0.5803135265134429,0.035906491399742724,0.5803135265134429,0.035906491399742724,**0.6164334312651423**,0.05865281211730978,,,0.5834655861681564,0.03612002100511133,0.16533245468793387,0.07204845491738118,0.31114121014873186,0.04318974648737592,**0.13643068556031035**,0.0202216302563364,0.17954962468723934,0.042108042550504994,5,MacroF1

MESSIDOR,0.05,flair,False,0.6467812300887237,0.04312831717607067,0.6467812300887237,0.04312831717607067,**0.6925411105457326**,0.026475916445255092,,,0.6484810497297366,0.043641978041704 9,0.2954127871235251,0.08605375860305729,0.20103504051764803,0.024391703756270273,**0.10226965255215632**,0.012398981224328396,0.27659716430358633,0.029001505810126 03,5,MacroF1

MESSIDOR,0.05,flair,True,0.6467812300887237,0.04312831717607067,0.6467812300887237,0.04312831717607067,**0.6925411105457326**,0.026475916445255092,,,0.6484810497297366,0.043641978041704 9,0.2954127871235251,0.08605375860305729,0.1501922427117824,0.02389986926055703,**0.0859207005741667**,0.00786555431494616,0.27659716430358633,0.029001505810126 03,5,MacroF1

MESSIDOR,0.05,openclip,False,0.6111743678738494,0.031288992450657235,0.6111743678738494,0.031288992450657235,**0.6706341358060667**,0.036857391990183064,,,0.6119651676847558,0.030683255822226304,0.22472402315396645,0.06237534072764541,0.10173489098747568,0.02453994615852942,**0.06722989160250714**,0.009998565693953297,0.24732276897414512,0.062847000278823353,5,MacroF1

MESSIDOR,0.05,openclip,True,0.6111743678738494,0.031288992450657235,0.6111743678738494,0.031288992450657235,**0.6706341358060667**,0.036857391990183064,,,0.6119651676847558,0.030683255822226304,0.22472402315396645,0.06237534072764541,0.06693955942988392,0.019583404046279878,**0.048089825364388425**,0.01047328674241922,0.24732276897414512,0.062847000278823353,5,MacroF1

MESSIDOR,0.05,retfound,False,0.5428724061905339,0.04641230456625338,0.5428724061905339,0.04641230456625338,**0.5872084194256615**,0.04620263708395139,,,0.5451119228834755,0.045785951924199784,0.09162670546077303,0.09314851650548757,0.28565539543827373,0.02145220

1230833357,**0.13293418549788721**,0.013852351361921545,0.15392827356130104,0.038339622255
83253,5,MacroF1
MESSIDOR,0.05,retfound,True,0.5428724061905339,0.04641230456625338,0.5428724061905339,
0.04641230456625338,**0.5872084194256615**,0.04620263708395139,,,0.5451119228834755,0.0457
85951924199784,0.09162670546077303,0.09314851650548757,0.22731726666291552,0.015551528
723129957,**0.11380653165240813**,0.008821642623295992,0.15392827356130104,0.0383396222558
3253,5,MacroF1
MESSIDOR,0.1,biomedclip,False,0.5962873671181391,0.028040919930766513,0.59628736711813
91,0.028040919930766513,**0.6494836923703791**,0.02505687918338518,,,0.5967669867774917,0.
027323337146806837,0.19399054732092527,0.05500446762191487,0.33844392950336133,0.02025
735593682144,**0.16912174844275754**,0.012246400933452278,0.18687239366138447,0.0291094775
12019335,5,MacroF1
MESSIDOR,0.1,biomedclip,True,0.5962873671181391,0.028040919930766513,0.596287367118139
1,0.028040919930766513,**0.6494627236339541**,0.02511979302513438,,,0.5967669867774917,0.0
27323337146806837,0.19399054732092527,0.05500446762191487,0.30241117954254143,0.023438
846839328648,**0.15001234141895028**,0.020451558583403605,0.18687239366138447,0.0291094775
12019335,5,MacroF1
MESSIDOR,0.1,flair,False,0.6749256917350969,0.029111537445303152,0.6749256917350969,0.
029111537445303152,**0.7459924315477225**,0.032200440970079985,,,0.675731493802087,0.02783
280908428533,0.3527643069006593,0.05660628245480312,0.1685669895509878,0.0192751609360
28466,**0.10756271852727448**,0.007143604114745336,0.326255212677231,0.10940186201258235,5
,MacroF1
MESSIDOR,0.1,flair,True,0.6749256917350969,0.029111537445303152,0.6749256917350969,0.0
29111537445303152,**0.7459924315477225**,0.032200440970079985,,,0.675731493802087,0.027832
80908428533,0.3527643069006593,0.05660628245480312,0.11538159171740207,0.0167068253553
60864,**0.07771844797894324**,0.005057292663795574,0.326255212677231,0.10940186201258235,5,
MacroF1
MESSIDOR,0.1,openclip,False,0.6236962377554542,0.0469573213258063,0.6236962377554542,0
.0469573213258063,**0.6833861120320958**,0.04068171152035707,,,0.6265362405645339,0.046752
583335831155,0.25398030563170176,0.09022499625020004,0.11582368776202201,0.02106722716
248168,**0.0804761083804385**,0.011957760874954516,0.25259382819015846,0.08243706054391223
,5,MacroF1
MESSIDOR,0.1,openclip,True,0.6236962377554542,0.0469573213258063,0.6236962377554542,0.
0469573213258063,**0.6833861120320958**,0.04068171152035707,,,0.6265362405645339,0.0467525
83335831155,0.25398030563170176,0.09022499625020004,0.06897272850076352,0.014487517995
874985,**0.05445937902763394**,0.005845584380389192,0.25259382819015846,0.0824370605439122
3,5,MacroF1
MESSIDOR,0.1,retfound,False,0.5787320036898963,0.03602105506818027,0.5787320036898963,
0.03602105506818027,**0.6221400148979112**,0.028703748098936707,,,0.58030687639974,0.03469
401731870863,0.16174845764035886,0.07021735196270362,0.23102998668948804,0.03676678703
331269,**0.11812029911561446**,0.013345814609730285,0.19964970809007504,0.0650559333404952
5,5,MacroF1

MESSIDOR,0.1,retfound,True,0.5787320036898963,0.03602105506818027,0.5787320036898963,0.03602105506818027,**0.6221400148979112**,0.028703748098936707,,,0.58030687639974,0.03469401731870863,0.16174845764035886,0.07021735196270362,0.17511916816234585,0.03729335965588428,**0.0953740129530177**,0.010377798785239917,0.19964970809007504,0.06505593334049525,5,MacroF1
MESSIDOR,0.2,biomedclip,False,0.6133541595268805,0.03787120761096594,0.6133541595268805,0.03787120761096594,**0.6688059574013773**,0.04831661371337593,,,0.6143672134294726,0.03734196266144305,0.2294203838269119,0.07519150453097197,0.3175643407801787,0.032227194133059016,**0.16570330482015574**,0.021088120770918440,0.25452877397831525,0.09277628125943924,5,MacroF1
MESSIDOR,0.2,biomedclip,True,0.6133541595268805,0.03787120761096594,0.6133541595268805,0.03787120761096594,**0.6687989541098303**,0.048297165348775685,,,0.6143672134294726,0.03734196266144305,0.2294203838269119,0.07519150453097197,0.2861301424602667,0.029634965599047658,**0.13738021900108943**,0.015405792434375492,0.25452877397831525,0.09277628125943924,5,MacroF1
MESSIDOR,0.2,flair,False,0.700198454623589,0.02460304469489394,0.700198454623589,0.02460304469489394,**0.7806108408014704**,0.02551796319979954,,,0.7006874489751442,0.02615593006642714,0.40126574316010205,0.049374314252086035,0.14201755175987876,0.031436876528397457,**0.09630921804254028**,0.01819615225080825,0.40083402835696413,0.08768133413118662,5,MacroF1
MESSIDOR,0.2,flair,True,0.700198454623589,0.02460304469489394,0.700198454623589,0.02460304469489394,**0.7806108408014704**,0.02551796319979954,,,0.7006874489751442,0.02615593006642714,0.40126574316010205,0.049374314252086035,0.08732600410779315,0.03352617712139063,**0.06583067875292202**,0.012520171734427894,0.40083402835696413,0.08768133413118662,5,MacroF1
MESSIDOR,0.2,openclip,False,0.6484212863049613,0.05141188106953432,0.6484212863049613,0.05141188106953432,**0.729051457541688**,0.05136527724448883,,,0.6492036620848163,0.050272090746135124,0.29954113129978266,0.09927887205938611,0.1123180045187473,0.046797638419492 63,**0.07804508358318034**,0.015287273642314340,0.30400333611342784,0.037191493146794 54,5,MacroF1
MESSIDOR,0.2,openclip,True,0.6484212863049613,0.05141188106953432,0.6484212863049613,0.05141188106953432,**0.729051457541688**,0.05136527724448883,,,0.6492036620848163,0.050272090746135124,0.29954113129978266,0.09927887205938611,0.07256403813759482,0.033656907068708795,**0.05964072682367072**,0.009534477279797707,0.30400333611342784,0.037191493146794 54,5,MacroF1
MESSIDOR,0.2,retfound,False,0.6193731776379771,0.024557805767334476,0.6193731776379771,0.024557805767334476,**0.668540716794236**,0.02269464222772422,,,0.6194279926950281,0.024371533454507547,0.23955707368946605,0.048283679036392795,0.1998361206054687,0.018404613108329693,**0.10385324913388135**,0.010506549094996227,0.2123603002502085,0.045363568273222252,5,MacroF1
MESSIDOR,0.2,retfound,True,0.6193731776379771,0.024557805767334476,0.6193731776379771,0.024557805767334476,**0.668540716794236**,0.02269464222772422,,,0.6194279926950281,0.0243

71533454507547,0.23955707368946605,0.048283679036392795,0.14264860605200 13,0.018995244
38192624,**0.08631647924382038**,0.010552183407130532,0.2123603002502085,0.045363568273222
52,5,MacroF1
ODIR200x3,0.05,biomedclip,False,0.8567506039731558,0.03865394099705653,0.8567506039731
558,0.03865394099705653,,,0.9669374999999999,0.011118199778716196,0.8566666666666667,0
.038819382329507025,0.7849999999999999,0.05822907349426057,0.09682298551003135,0.02373
1465305086664,**0.15887981448019078**,0.025102132707245036,,,5,MacroF1
ODIR200x3,0.05,biomedclip,True,0.8567506039731558,0.03865394099705653,0.85675060397315
58,0.03865394099705653,,,0.9673124999999999,0.011022339610360012,0.8566666666666667,0.
038819382329507025,0.7849999999999999,0.05822907349426057,0.05942434216539062,0.017347
59712962752,**0.1350572944918148**,0.01703680984519239,,,5,MacroF1
ODIR200x3,0.05,flair,False,0.8432129936017612,0.04329802372403759,0.8432129936017612,0
.04329802372403759,,,0.9632708333333333,0.012635249208838092,0.8433333333333334,0.0418
3300132670376,0.765,0.06274950199005568,0.0730249109367529,0.014692414082001616,**0.1527
050924101106**,0.028999857672126705,,,5,MacroF1
ODIR200x3,0.05,flair,True,0.8432129936017612,0.04329802372403759,0.8432129936017612,0.
04329802372403759,,,0.9630624999999998,0.012802460497563223,0.8433333333333334,0.04183
300132670376,0.765,0.06274950199005568,0.11610556095838542,0.027209028726703173,**0.1300
8147405020126**,0.012336753038959467,,,5,MacroF1
ODIR200x3,0.05,openclip,False,0.7928908654087643,0.05633097714063105,0.792890865408764
3,0.05633097714063105,,,0.937375,0.016888240690229095,0.795,0.053877432917482065,0.692
5000000000001,0.08081614937622307,0.10529152005910869,0.020801434022754434,**0.136558449
87148616**,0.012063212718857927,,,5,MacroF1
ODIR200x3,0.05,openclip,True,0.7928908654087643,0.05633097714063105,0.7928908654087643
,0.05633097714063105,,,0.9374375,0.016865126998648918,0.795,0.053877432917482065,0.692
5000000000001,0.08081614937622307,0.13792176748315488,0.03357574038624564,**0.1195177897
9916568**,0.02190644573655521,,,5,MacroF1
ODIR200x3,0.05,retfound,False,0.6497396626644518,0.0458080485070679,0.6497396626644518
,0.0458080485070679,,,0.888649741255008,0.013870553944312532,0.5862515618924804,0.0395
21898323069994,0.4879311417937172,0.05458671599432141,0.04249455112281056,0.0095243480
98457168,**0.11372345830422363**,0.014946556332433315,,,5,MacroF1
ODIR200x3,0.05,retfound,True,0.6497396626644518,0.0458080485070679,0.6497396626644518,
0.0458080485070679,,,0.8884365238511075,0.013514535873743502,0.5862515618924804,0.0395
21898323069994,0.4879311417937172,0.05458671599432141,0.03227237829124782,0.0110864600
39895888,**0.08122189227877022**,0.01724591140008306,,,5,MacroF1
ODIR200x3,0.1,biomedclip,False,0.8924159658489422,0.036155046684827316,0.8924159658489
422,0.036155046684827316,,,0.9747708333333334,0.01122136595158739,0.8916666666666668,0
.03679900360969934,0.8375,0.05519850541454906,0.07726348146796222,0.020860660392918905
,**0.1601501638780016**,0.05321836010662043,,,5,MacroF1
ODIR200x3,0.1,biomedclip,True,0.8924159658489422,0.036155046684827316,0.89241596584894
22,0.036155046684827316,,,0.9746666666666666,0.011572443172131902,0.8916666666666668,0

.036799003609699 34,0.8375,0.05519850541454906,0.06466856280962621,0.017467021358802085,**0.1493700674688015**,0.01737734659825441,,,5,MacroF1
ODIR200x3,0.1,flair,False,0.8731988676019956,0.03167645685824137,0.8731988676019956,0.03167645685824137,,,0.9711041666666667,0.010202113326142044,0.8733333333333334,0.030844592538869644,0.8099999999999999,0.04626688880830437,0.06463112617532409,0.008352715378148001,**0.17627075487686011**,0.02617252583879942,,,5,MacroF1
ODIR200x3,0.1,flair,True,0.8731988676019956,0.03167645685824137,0.8731988676019956,0.03167645685824137,,,0.9713750000000001,0.009983004220368248,0.8733333333333334,0.030844592538869644,0.8099999999999999,0.04626688880830437,0.09373360017935432,0.022411049892035975,**0.14942147058013341**,0.027194472143958464,,,5,MacroF1
ODIR200x3,0.1,openclip,False,0.8707491905623256,0.0386797044179574,0.8707491905623256,0.0386797044179574,,,0.9666458333333333,0.015388558272120 39,0.8716666666666667,0.039352396510391 91,0.8074999999999999,0.05902859476558797,0.09084115942319228,0.022950634817324077,**0.1486780032951402**,0.030946704195189866,,,5,MacroF1
ODIR200x3,0.1,openclip,True,0.8707491905623256,0.0386797044179574,0.8707491905623256,0.0386797044179574,,,0.9665833333333333,0.015060353120398974,0.8716666666666667,0.039352396510391 91,0.8074999999999999,0.05902859476558797,0.15824262370665865,0.034840171873293 07,**0.14145708631408477**,0.022953633831338357,,,5,MacroF1
ODIR200x3,0.1,retfound,False,0.7442519846496616,0.010295320315371038,0.7442519846496616,0.010295320315371038,,,0.9392141341716325,0.012876230937106228,0.6855477150138695,0.012640710744144036,0.625346437496646,0.01601842193177645,0.025031564352810582,0.0048981860085942274,**0.11428400170636199**,0.0131455301755216,,,5,MacroF1
ODIR200x3,0.1,retfound,True,0.7442519846496616,0.010295320315371038,0.7442519846496616,0.010295320315371038,,,0.939172476839874,0.012746472592929537,0.6855477150138695,0.012640710744144036,0.625346437496646,0.01601842193177645,0.055478274795608594,0.005315385567592085,**0.0975710740400202**,0.008175915832669196,,,5,MacroF1
ODIR200x3,0.2,biomedclip,False,0.8701949732906084,0.024487106035019513,0.8701949732906084,0.024487106035019513,,,0.97,0.008933468399165525,0.8699999999999999,0.024008100484813173,0.805,0.036012150727219826,0.10429594536622362,0.025357812873199558,**0.17632933496914133**,0.027254259069381 02,,,5,MacroF1
ODIR200x3,0.2,biomedclip,True,0.8701949732906084,0.024487106035019513,0.8701949732906084,0.024487106035019513,,,0.9701041666666667,0.00897920291949369,0.8699999999999999,0.024008100484813173,0.805,0.036012150727219826,0.07667355934778843,0.019915032119093404,**0.17027346948399982**,0.02855957660364089,,,5,MacroF1
ODIR200x3,0.2,flair,False,0.8999962235899052,0.051764262349825535,0.8999962235899052,0.051764262349825535,,,0.9769583333333334,0.01384797794104084,0.9,0.051706973524961876,0.85,0.07756046028744289,0.06507209683458004,0.016566540823524534,**0.1577192474287012**,0.03361959531397522,,,5,MacroF1
ODIR200x3,0.2,flair,True,0.8999962235899052,0.051764262349825535,0.8999962235899052,0.051764262349825535,,,0.9772083333333332,0.013626815628262796,0.9,0.051706973524961876,0.85,0.07756046028744289,0.07855771422386164,0.029945071889554357,**0.1498634979425768**,0.039572093837358976,,,5,MacroF1

```
ODIR200x3,0.2,openclip,False,0.8781483280606208,0.027348480257930182,0.8781483280606208,0.027348480257930182,,,0.9690208333333332,0.008958817816356587,0.8783333333333333,0.026744677559801984,0.8175000000000001,0.04011701633970304,0.0733096659680207,0.014586829733662354,0.16650623187230013,0.01680465492634328,,,5,MacroF1
ODIR200x3,0.2,openclip,True,0.8781483280606208,0.027348480257930182,0.8781483280606208,0.027348480257930182,,,0.9692291666666666,0.00865417519258137,0.8783333333333333,0.026744677559801984,0.8175000000000001,0.04011701633970304,0.11399430269996322,0.016377480785352755,0.14695230601363413,0.01556616799556443,,,5,MacroF1
ODIR200x3,0.2,retfound,False,0.8201449400037589,0.02946175979205535,0.8201449400037589,0.02946175979205535,,,0.9658468660319619,0.008491460783608257,0.7685202070417362,0.02141136404030205,0.731387546655164,0.045785631140617004,0.021243495773348962,0.011046997636922566,0.08905054447702351,0.020074216574594867,,,5,MacroF1
ODIR200x3,0.2,retfound,True,0.8201449400037589,0.02946175979205535,0.8201449400037589,0.02946175979205535,,,0.9658077017292779,0.008598967706405879,0.7685202070417362,0.02141136404030205,0.731387546655164,0.045785631140617004,0.058188953019739184,0.018577476380185358,0.08104275542962511,0.022746543844680077,,,5,MacroF1
REFUGE,0.05,biomedclip,False,0.756076388888889,0.07227572554120587,0.5957677454575984,0.08838599192621151,0.756076388888889,0.07227572554120587,,,0.5708333333333334,0.055858548756083926,0.21328671328671328,0.16901964630829988,0.08903201550245284,0.014118459096327201,0.1324496170092,0.018898186198306598,0.23055555555555554,0.19681099537687055,5,AUROC
REFUGE,0.05,biomedclip,True,0.7559027777777778,0.07257069044650685,0.5957677454575984,0.08838599192621151,0.7559027777777778,0.07257069044650685,,,0.5708333333333334,0.055858548756083926,0.21328671328671328,0.16901964630829988,0.08133716762065883,0.013667427477987241,0.12992302954950388,0.01806753051537269,0.23611111111111108,0.18890931262132563,5,AUROC
REFUGE,0.05,flair,False,0.7180555555555556,0.13761131156626502,0.47368421052631576,0.0,0.7180555555555556,0.13761131156626502,,,0.5,0.0,0.0,0.0,0.08482131496071807,0.008629681706778476,0.1449997320964973,0.05330556366656239,0.28888888888888886,0.2915674434398911,5,AUROC
REFUGE,0.05,flair,True,0.7180555555555556,0.13761131156626502,0.47368421052631576,0.0,0.7180555555555556,0.13761131156626502,,,0.5,0.0,0.0,0.0,0.07544615089893338,0.007447623008566778,0.13239275317140947,0.04728838525429 91,0.28888888888888886,0.2915674434398911,5,AUROC
REFUGE,0.05,openclip,False,0.8065972222222222,0.09896290194132423,0.47368421052631576,0.0,0.8065972222222222,0.09896290194132423,,,0.5,0.0,0.0,0.0,0.07335923060774802,0.01167640777667559,0.083222869108567,0.024326659861363155,0.43611111111111106,0.29439858133380675,5,AUROC
REFUGE,0.05,openclip,True,0.8065972222222222,0.09896290194132423,0.47368421052631576,0.0,0.8065972222222222,0.09896290194132423,,,0.5,0.0,0.0,0.0,0.05083685576915736,0.016678725935114887,0.08628903538908637,0.016017918057241946,0.43611111111111106,0.29439858133380675,5,AUROC
```

REFUGE,0.05,retfound,False,0.663888888888889,0.14927830016087318,0.49371732960711173,0.05284515572226629,**0.663888888888889**,0.14927830016087318,,,0.5069444444444444,0.03220006422047119,0.02482517482517482,0.10320165665117478,0.09203641623258589,0.0074640255653127124,**0.11817375064747968**,0.0546840367291041,0.225,0.16298006013006622,5,AUROC
REFUGE,0.05,retfound,True,0.663888888888889,0.14927830016087318,0.49371732960711173,0.05284515572226629,**0.663888888888889**,0.14927830016087318,,,0.5069444444444444,0.03220006422047119,0.02482517482517482,0.10320165665117478,0.07661161854863162,0.012152387184881405,**0.10850501506440188**,0.027221146863893916,0.225,0.16298006013006622,5,AUROC
REFUGE,0.1,biomedclip,False,0.8079861111111111,0.0808163917968581,0.6130801787586199,0.09236628625877662,**0.8079861111111111**,0.0808163917968581,,,0.601388888888889,0.08166737528037243,0.24518259518259516,0.16445050964897298,0.09109575793147087,0.013999930023233915,**0.15677089302560815**,0.10074953447591926,0.25277777777777777,0.32295773335045597,5,AUROC
REFUGE,0.1,biomedclip,True,0.8086805555555555,0.08136278748723458,0.6130801787586199,0.09236628625877662,**0.8086805555555555**,0.08136278748723458,,,0.601388888888889,0.08166737528037243,0.24518259518259516,0.16445050964897298,0.08551963374018665,0.013946166416860578,**0.1591590589730289**,0.06798221381972208,0.30833333333333335,0.2889089202456907,5,AUROC
REFUGE,0.1,flair,False,0.8427083333333334,0.10409140799862307,0.47368421052631576,0.0,**0.8427083333333334**,0.10409140799862307,,,0.5,0.0,0.0,0.0,0.08566822439432142,0.006549016745764281,**0.1480202053618262**,0.06349687554700359,0.5361111111111111,0.28406088505768107,5,AUROC
REFUGE,0.1,flair,True,0.8427083333333334,0.10409140799862307,0.47368421052631576,0.0,**0.8427083333333334**,0.10409140799862307,,,0.5,0.0,0.0,0.0,0.06630906239151953,0.009240292905343157,**0.13725263003871807**,0.04804145465121094,0.5361111111111111,0.28406088505768107,5,AUROC
REFUGE,0.1,openclip,False,0.8534722222222222,0.09657366764899739,0.5422330661089811,0.06257625748964958,**0.8534722222222222**,0.09657366764899739,,,0.5375,0.03423265984407287,0.12272727272727271,0.11203415948969304,0.07855358988046643,0.007043021615274124,**0.17296764186893157**,0.05451665760654054,0.5333333333333333,0.3131320768156214,5,AUROC
REFUGE,0.1,openclip,True,0.8534722222222222,0.09657366764899739,0.5422330661089811,0.06257625748964958,**0.8534722222222222**,0.09657366764899739,,,0.5375,0.03423265984407287,0.12272727272727271,0.11203415948969304,0.06081725701689718,0.014705093903854516,**0.15421867284534568**,0.045262625103599415,0.5333333333333333,0.3131320768156214,5,AUROC
REFUGE,0.1,retfound,False,0.810763888888889,0.06538711661550858,0.5791729212656366,0.07584739853762092,**0.810763888888889**,0.06538711661550858,,,0.5583333333333333,0.04925173118471429,0.18181818181818177,0.14876770368678396,0.07888400360941886,0.015034906017653457,**0.13167397998609104**,0.03395039033770607,0.5,0.25,5,AUROC
REFUGE,0.1,retfound,True,0.810763888888889,0.06538711661550858,0.5791729212656366,0.07584739853762092,**0.810763888888889**,0.06538711661550858,,,0.5583333333333333,0.04925173118471429,0.18181818181818177,0.14876770368678396,0.0657543101906776,0.008139361273513373,**0.11897660237710314**,0.03172842570455016,0.5,0.25,5,AUROC

```
REFUGE,0.2,biomedclip,False,0.8743055555555556,0.06930976523621162,0.6635091596012461,
0.11297629982132087,0.8743055555555556,0.06930976523621162,,,0.6402777777777777,0.0881
6983471168825,0.3417542016806723,0.20544842987029557,0.0791813847422599,0.014998782494
736074,0.1150279029021132,0.09498613608316296,0.5638888888888889,0.23128232173185875,5
,AUROC
REFUGE,0.2,biomedclip,True,0.8743055555555556,0.06930976523621164,0.6635091596012461,0
.11297629982132087,0.8743055555555556,0.06930976523621164,,,0.6402777777777777,0.08816
983471168825,0.3417542016806723,0.20544842987029557,0.07639881342649456,0.018094580769
820748,0.12869917521603585,0.08819864180536863,0.5638888888888889,0.23128232173185875,
5,AUROC
REFUGE,0.2,flair,False,0.8704861111111111,0.07274699173916926,0.6351008491935645,0.046
10371886860245,0.8704861111111111,0.07274699173916926,,,0.5972222222222222,0.032200064
220471204,0.288961038961039,0.08524659644880601,0.07246439725160596,0.0089482204110940
7,0.16564304986641545,0.0653863241253936,0.5527777777777778,0.24775224083649816,5,AURO
C
REFUGE,0.2,flair,True,0.8704861111111111,0.07274699173916926,0.6351008491935645,0.0461
0371886860245,0.8704861111111111,0.07274699173916926,,,0.5972222222222222,0.0322000642
20471204,0.288961038961039,0.08524659644880601,0.06330432459712022,0.00677877771550348
34,0.15240934267142606,0.040891576045867704,0.5527777777777778,0.24775224083649816,5,A
UROC
REFUGE,0.2,openclip,False,0.890625,0.06479672780042195,0.5869957505111161,0.1219622000
2544016,0.890625,0.06479672780042195,,,0.5722222222222222,0.08077675568664082,0.202564
10256410257,0.2231008681519443,0.07111101239919657,0.014001495617078932,0.173082397924
10215,0.05434310475826862,0.611111111111111,0.24825781849610387,5,AUROC
REFUGE,0.2,openclip,True,0.890625,0.06479672780042195,0.5869957505111161,0.12196220002
544016,0.890625,0.06479672780042195,,,0.5722222222222222,0.08077675568664082,0.2025641
0256410257,0.2231008681519443,0.06150141879916184,0.011459309733646594,0.1453332056562
127,0.041065394876489233,0.611111111111111,0.24825781849610387,5,AUROC
REFUGE,0.2,retfound,False,0.835763888888889,0.05719425641461728,0.6606320663361421,0.0
6250401270355249,0.835763888888889,0.05719425641461728,,,0.6194444444444445,0.04484412
3030452605,0.33404095904095904,0.11980497613685914,0.07193907693028444,0.0140496981867
35876,0.11955776595321602,0.04576935376073647,0.55,0.1118033988749895,5,AUROC
REFUGE,0.2,retfound,True,0.835763888888889,0.05719425641461728,0.6606320663361421,0.06
250401270355249,0.835763888888889,0.05719425641461728,,,0.6194444444444445,0.044844123
030452605,0.33404095904095904,0.11980497613685914,0.058034053891897176,0.0129897762867
58284,0.124770984294967,0.026911356691108352,0.55,0.1118033988749895,5,AUROC
```

Table A.1. *Primary metric (mean [95% CI]) across datasets, models, and label fractions.*

Macro-F1 for MESSIDOR and ODIR-200×3; AUROC for REFUGE; results reported with and without temperature scaling (calibrated = True/False), averaged over 5 folds.

Primary (mean [95% CI])

dataset	model	frac	calibrated	PrimaryMetric	ci_str	folds
MESSIDOR	flair	0.05	False	MacroF1	0.647 [0.593,0.700]	5
MESSIDOR	openclip	0.05	False	MacroF1	0.611 [0.572,0.650]	5
MESSIDOR	biomedclip	0.05	False	MacroF1	0.580 [0.536,0.625]	5
MESSIDOR	retfound	0.05	False	MacroF1	0.543 [0.485,0.601]	5
MESSIDOR	flair	0.05	True	MacroF1	0.647 [0.593,0.700]	5
MESSIDOR	openclip	0.05	True	MacroF1	0.611 [0.572,0.650]	5
MESSIDOR	biomedclip	0.05	True	MacroF1	0.580 [0.536,0.625]	5
MESSIDOR	retfound	0.05	True	MacroF1	0.543 [0.485,0.601]	5
MESSIDOR	flair	0.1	False	MacroF1	0.675 [0.639,0.711]	5
MESSIDOR	openclip	0.1	False	MacroF1	0.624 [0.565,0.682]	5
MESSIDOR	biomedclip	0.1	False	MacroF1	0.596 [0.561,0.631]	5
MESSIDOR	retfound	0.1	False	MacroF1	0.579 [0.534,0.623]	5
MESSIDOR	flair	0.1	True	MacroF1	0.675 [0.639,0.711]	5
MESSIDOR	openclip	0.1	True	MacroF1	0.624 [0.565,0.682]	5
MESSIDOR	biomedclip	0.1	True	MacroF1	0.596 [0.561,0.631]	5
MESSIDOR	retfound	0.1	True	MacroF1	0.579 [0.534,0.623]	5

MESSIDOR	flair	0.2	False	MacroF1	0.700 [0.670,0.731]	5
MESSIDOR	openclip	0.2	False	MacroF1	0.648 [0.585,0.712]	5
MESSIDOR	retfound	0.2	False	MacroF1	0.619 [0.589,0.650]	5
MESSIDOR	biomedclip	0.2	False	MacroF1	0.613 [0.566,0.660]	5
MESSIDOR	flair	0.2	True	MacroF1	0.700 [0.670,0.731]	5
MESSIDOR	openclip	0.2	True	MacroF1	0.648 [0.585,0.712]	5
MESSIDOR	retfound	0.2	True	MacroF1	0.619 [0.589,0.650]	5
MESSIDOR	biomedclip	0.2	True	MacroF1	0.613 [0.566,0.660]	5
ODIR200x3	biomedclip	0.05	False	MacroF1	0.857 [0.809,0.905]	5
ODIR200x3	flair	0.05	False	MacroF1	0.843 [0.789,0.897]	5
ODIR200x3	openclip	0.05	False	MacroF1	0.793 [0.723,0.863]	5
ODIR200x3	retfound	0.05	False	MacroF1	0.650 [0.593,0.707]	5
ODIR200x3	biomedclip	0.05	True	MacroF1	0.857 [0.809,0.905]	5
ODIR200x3	flair	0.05	True	MacroF1	0.843 [0.789,0.897]	5
ODIR200x3	openclip	0.05	True	MacroF1	0.793 [0.723,0.863]	5
ODIR200x3	retfound	0.05	True	MacroF1	0.650 [0.593,0.707]	5
ODIR200x3	biomedclip	0.1	False	MacroF1	0.892 [0.848,0.937]	5
ODIR200x3	flair	0.1	False	MacroF1	0.873 [0.834,0.913]	5
ODIR200x3	openclip	0.1	False	MacroF1	0.871 [0.823,0.919]	5
ODIR200x3	retfound	0.1	False	MacroF1	0.744 [0.731,0.757]	5
ODIR200x3	biomedclip	0.1	True	MacroF1	0.892 [0.848,0.937]	5

ODIR200x3	flair	0.1	True	MacroF1	0.873 [0.834,0.913]	5
ODIR200x3	openclip	0.1	True	MacroF1	0.871 [0.823,0.919]	5
ODIR200x3	retfound	0.1	True	MacroF1	0.744 [0.731,0.757]	5
ODIR200x3	flair	0.2	False	MacroF1	0.900 [0.836,0.964]	5
ODIR200x3	openclip	0.2	False	MacroF1	0.878 [0.844,0.912]	5
ODIR200x3	biomedclip	0.2	False	MacroF1	0.870 [0.840,0.901]	5
ODIR200x3	retfound	0.2	False	MacroF1	0.820 [0.784,0.857]	5
ODIR200x3	flair	0.2	True	MacroF1	0.900 [0.836,0.964]	5
ODIR200x3	openclip	0.2	True	MacroF1	0.878 [0.844,0.912]	5
ODIR200x3	biomedclip	0.2	True	MacroF1	0.870 [0.840,0.901]	5
ODIR200x3	retfound	0.2	True	MacroF1	0.820 [0.784,0.857]	5
REFUGE	openclip	0.05	False	AUROC	0.807 [0.684,0.929]	5
REFUGE	biomedclip	0.05	False	AUROC	0.756 [0.666,0.846]	5
REFUGE	flair	0.05	False	AUROC	0.718 [0.547,0.889]	5
REFUGE	retfound	0.05	False	AUROC	0.664 [0.479,0.849]	5
REFUGE	openclip	0.05	True	AUROC	0.807 [0.684,0.929]	5
REFUGE	biomedclip	0.05	True	AUROC	0.756 [0.666,0.846]	5
REFUGE	flair	0.05	True	AUROC	0.718 [0.547,0.889]	5
REFUGE	retfound	0.05	True	AUROC	0.664 [0.479,0.849]	5
REFUGE	openclip	0.1	False	AUROC	0.853 [0.734,0.973]	5
REFUGE	flair	0.1	False	AUROC	0.843 [0.713,0.972]	5

REFUGE	retfound	0.1	False	AUROC	0.811 [0.730,0.892]	5
REFUGE	biomedclip	0.1	False	AUROC	0.808 [0.708,0.908]	5
REFUGE	openclip	0.1	True	AUROC	0.853 [0.734,0.973]	5
REFUGE	flair	0.1	True	AUROC	0.843 [0.713,0.972]	5
REFUGE	retfound	0.1	True	AUROC	0.811 [0.730,0.892]	5
REFUGE	biomedclip	0.1	True	AUROC	0.809 [0.708,0.910]	5
REFUGE	openclip	0.2	False	AUROC	0.891 [0.810,0.971]	5
REFUGE	biomedclip	0.2	False	AUROC	0.874 [0.788,0.960]	5
REFUGE	flair	0.2	False	AUROC	0.870 [0.780,0.961]	5
REFUGE	retfound	0.2	False	AUROC	0.836 [0.765,0.907]	5
REFUGE	openclip	0.2	True	AUROC	0.891 [0.810,0.971]	5
REFUGE	biomedclip	0.2	True	AUROC	0.874 [0.788,0.960]	5
REFUGE	flair	0.2	True	AUROC	0.870 [0.780,0.961]	5
REFUGE	retfound	0.2	True	AUROC	0.836 [0.765,0.907]	5

Table C.1. *REFUGE per-fraction AUROC and sensitivity at 90% specificity (mean ± SD across folds) for each model, with/without calibration.*

Pairwise significance (winner vs runner-up)

dataset	frac	calibrated	winner	runner_up	primary_metric	mean_diff	ci95_lo	ci95_hi	p_value	folds
MESSIDOR	0.05	False	flair	openclip	MacroF1	0.0356069	-0.00145607	0.0726698	0.0559591	5
MESSIDOR	0.05	True	flair	openclip	MacroF1	0.0356069	-0.00145607	0.0726698	0.0559591	5

MESSIDOR	0.1	False	flair	openclip	MacroF1	0.0512295	-0.0167597	0.119219	0.10459	5
MESSIDOR	0.1	True	flair	openclip	MacroF1	0.0512295	-0.0167597	0.119219	0.10459	5
MESSIDOR	0.2	False	flair	openclip	MacroF1	0.0517772	-0.00621127	0.109766	0.0682803	5
MESSIDOR	0.2	True	flair	openclip	MacroF1	0.0517772	-0.00621127	0.109766	0.0682803	5
ODIR200x3	0.05	False	biomedclip	flair	MacroF1	0.0135376	-0.0330995	0.0601747	0.465464	5
ODIR200x3	0.05	True	biomedclip	flair	MacroF1	0.0135376	-0.0330995	0.0601747	0.465464	5
ODIR200x3	0.1	False	biomedclip	flair	MacroF1	0.0192171	-0.028981	0.0674152	0.330383	5
ODIR200x3	0.1	True	biomedclip	flair	MacroF1	0.0192171	-0.028981	0.0674152	0.330383	5
ODIR200x3	0.2	False	flair	openclip	MacroF1	0.0218479	-0.0358744	0.0795702	0.352607	5
ODIR200x3	0.2	True	flair	openclip	MacroF1	0.0218479	-0.0358744	0.0795702	0.352607	5
REFUGE	0.05	False	openclip	biomedclip	AUROC	0.0505208	-0.070397	0.171439	0.310558	5
REFUGE	0.05	True	openclip	biomedclip	AUROC	0.0506944	-0.0707559	0.172145	0.310965	5
REFUGE	0.1	False	openclip	flair	AUROC	0.0107639	-0.1	0.121528	0.800654	5
REFUGE	0.1	True	openclip	flair	AUROC	0.0107639	-0.1	0.121528	0.800654	5
REFUGE	0.2	False	openclip	biomedclip	AUROC	0.0163194	-0.107088	0.139726	0.7321	5
REFUGE	0.2	True	openclip	biomedclip	AUROC	0.0163194	-0.107088	0.139726	0.7321	5

Table D.1. *Pairwise comparison between the top two models per dataset×label-fraction (primary metric): mean difference, 95% CI, Wilcoxon p-value (Holm-adjusted), and number of folds. (Insert your "Pairwise significance (winner vs runner-up)" table here.)*
Caption note: "Positive mean_diff favors the winner; non-significant results (p≥0.05) are italicized."