

A Fine-Grained Benchmark for Ophthalmic Foundation Models with Point-Based Classification and Retrieval

Hyunmin Kim

University College London

Michael Grün

St. Franziskus Hospital

Rahul A. Jonas

University of Cologne

Heesuk Kim

Severance Hospital

Sungyeon Kim

Amazon (United States)

Takahiro Ninomiya

Tohoku University

Broder Poschkamp

Universität Greifswald

Xiaolei Diao

University College London

Sophie-Christin Ernst

University of Birmingham

Josef Huemer

Moorfields Eye Hospital NHS Foundation Trust

Koichi Ito

Tohoku University

Paul Nderitu

Moorfields Eye Hospital NHS Foundation Trust

Kuang Hu

Moorfields Eye Hospital NHS Foundation Trust

Lie Ju

University College London

Yiqun Lin

University College London

Yilan Wu

University College London

Ariel Yuhan Ong

University College London

Michael G. Morley

Massachusetts Eye and Ear Infirmary

Tobias Elze

Massachusetts Eye and Ear Infirmary

Junseo Choi

Hanyang University Guri Hospital

Danbi Lee

Hanyang University Guri Hospital

Youngbin Seo

Severance Hospital

Yaxing Wang

Tsinghua University

Songhomitra Panda-Jonas

Fondation de Rothschild

Jost B. Jonas

Fondation de Rothschild

Mukharram M. Bikbov

Ufa Eye Research Institute

Gyulli M. Kazakbaeva

Ufa Eye Research Institute

Vinay Nangia

Suraj Eye Institute

Clemens Lange

St. Franziskus Hospital

Albrecht Lommatzsch

St. Franziskus Hospital

Anthony P. Khawaja

University College London

Robert Luben

University College London

Alasdair Warwick

University College London

Andre Altmann

University College London

Alastair K. Denniston

University of Birmingham

Daniel C. Alexander

University College London

Siegfried K. Wagner

University College London

Yukun Zhou

University College London

Sunjin Hwang

Hanyang University Guri Hospital

Pearse A. Keane

p.keane@ucl.ac.uk

University College London

Article

Keywords:

Posted Date: September 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10817915/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: Competing interest reported. P.A.K. is a co-founder of Cascader Ltd. and has acted as a consultant for Skleo Health, insitro, Retina Consultants of America, Roche, Boehringer-Ingelheim, and Bitfount, and is an equity owner in Big Picture Medical. He has received speaker fees from Zeiss, Théa, Apellis, and Roche, travel support from Bayer and Roche, and grant funding from Roche. He has attended advisory boards for Topcon, Bayer, Boehringer-Ingelheim, and Roche. J.B.J. holds European patents EP 3 271 392 and WO 2025/056713 A1, Japanese patent JP 2021-119187, and U.S. Patent No. US 12,024,557 B2. J.H. has received travel support from Bayer and Roche and speaker fees from Zeiss, Roche, and Bayer. S.P.J. holds European patent EP 3 271 392, Japanese patent JP 2021-119187, and U.S. Patent No. US 12,024,557 B2: Agents for use in the therapeutic or prophylactic treatment of myopia or hyperopia; and European patent application WO 2025/056713 A1: EGFR Antagonists for the treatment of diseases involving unwanted migration, proliferation, and metaplasia of retinal pigment epithelium (RPE) cells. A.P.K has acted as a paid consultant or lecturer to Abbvie, Aerie, Google Health, Heidelberg Engineering, Glaucore, Novartis, Qlaris Bio, Regeneron, Reichert, Santen, Seonix Bio, Tavo Biotherapeutics, Thea and Topcon. The remaining authors declare no competing interests.

Abstract

General-purpose vision foundation models have been evaluated across a wide range of tasks to demonstrate the versatility of their learned representations. By contrast, ophthalmic foundation models have largely been assessed using a narrow set of downstream tasks, primarily disease classification and segmentation, providing only a limited view of their representational capabilities. Considering how clinicians interpret ophthalmic images, where disease-relevant clinical signs are identified before diagnosis, we introduced an ophthalmic foundation model evaluation framework comprising two tasks: 1) point-supervised disease classification and 2) point-guided clinical sign-based image retrieval. Each task incorporates sparse point annotations indicating disease-relevant evidence locations. We applied this approach to diabetic retinopathy using colour fundus photography (CFP) and age-related macular degeneration using both CFP and optical coherence tomography (OCT) scans. Our evaluation spanned 11 foundation models across four newly curated and six established datasets. The results reveal that existing ophthalmic foundation models, despite their strong performance on disease classification, still struggle with the clinical sign retrieval task, although they outperform general-purpose foundation models. Our findings highlight the need to improve these models' capabilities at the fine-grained sign-level. We publicly release our benchmark to facilitate further research on ophthalmic foundation models.

Introduction

General-purpose vision foundation models have advanced rapidly, from the DINO series^{1–3} to more recent encoders, such as Perception Encoder⁴ and the latest Efficient Universal Perception Encoder,⁵ all aiming to learn universal representations. Their generalisability has been validated across a wide range of downstream tasks, including object tracking, semantic correspondence, and visual question–answers, reflecting the expectation that a truly capable foundation model should be generalisable across diverse visual understanding tasks. By contrast, evaluations of ophthalmic foundation models, such as RETFound,^{6,7} UrFound,⁸ OCT-CUBE,⁹ and RetiZero,¹⁰ have mainly focused on disease classification. More recently, evaluations have extended beyond classification. RetiZero adopted disease-level image retrieval as a clinically relevant benchmark for identifying similar images within large, unlabelled datasets.¹⁰ Simultaneously, the optical coherence tomography (OCT) foundation model MIRAGE¹¹ was evaluated for lesion and layer segmentation in OCT, which required an additional segmentation head. Nevertheless, most existing evaluations focus on classification and fail to fully reflect how clinicians interpret ophthalmic images.

Clinical decision-making is based on the recognition and integration of clinically meaningful signs, rather than on raw pixels alone; clinicians interpret these signs to arrive at a diagnosis or clinical decision. For example, the grade of diabetic retinopathy (DR) in colour fundus photographs (CFP) is defined by multiple localised clinical signs, such as retinal haemorrhages and cotton wool spots,^{12,13} whereas age-related macular degeneration (AMD) requiring intravitreal anti-vascular endothelial growth factor treatment is characterised by the presence of intraretinal or subretinal fluid in OCT.^{14,15} To better reflect this diagnostic process, previous studies^{16,17} have adopted two-stage approaches that explicitly model clinical signs through segmentation or lesion detection before disease classification. Despite these efforts, clinical signs remain largely overlooked in the evaluation of ophthalmic foundation models, which continue to focus on disease-level classification.^{6,18}

To address this gap, we propose a fine-grained evaluation framework based on point annotations that indicate the locations of diagnostically relevant clinical signs. This framework comprises two tasks: point-supervised disease classification and point-guided clinical sign-based image retrieval. The clinical sign retrieval task, which is one of the key evaluation tasks for assessing a model's recognition and understanding of images in computer vision, evaluates whether a foundation model correctly recognises a given clinical sign and retrieves other cases that exhibit the same sign. This parallels an important component of clinical reasoning, in which clinicians recognise characteristic imaging signs, relate them to previously encountered patterns or cases, and integrate this information to reach a clinical interpretation.

Using this framework, we evaluated eight ophthalmic and three general-purpose foundation models across two imaging modalities (CFP for DR and AMD and OCT for AMD) on four newly curated datasets and six established datasets, including the UK Biobank,¹⁹ Ural Eye and Medical Study,²⁰ and Beijing Eye Study.²¹ This evaluation addressed two central questions: (1) Do general-purpose foundation models generalise to ophthalmic imaging? and (2) does strong discriminative performance translate into strong performance on retrieval, one measure of image understanding? Further, we examined a related sub-question: (3) Do point annotations provide a meaningful signal for disease classification fine-tuning and clinical sign retrieval? An overview of the study, including the datasets, models, and two evaluation tasks, is provided in Fig. 1.

Results

Overview

In this study, we focused on CFP and OCT, which together account for the majority of ophthalmic imaging and are acquired routinely in clinical practice. We targeted DR and AMD, two leading causes of blindness worldwide. Within the proposed evaluation framework, we evaluated 11 models: five CFP-based foundation models (RETFound-DINOv2, RETFound-Green,²² VisionFM-CFP,²³ RetiZero, and UrFound), three OCT slice-based foundation models (RETFound-MAE, VisionFM-OCT, and MIRAGE), and three general-purpose vision foundation models (DINOv3-Large, Perception Encoder-Large (PE-Large), and Efficient Universal Perception Encoder-Base (EUPE-Base)). For the CFP-based ophthalmic and general-purpose vision foundation models, DR and AMD classification performance was evaluated on the CFP modality. For the OCT slice-based ophthalmic and general-purpose vision foundation models, AMD classification was evaluated on the OCT modality. In Task 1, we swept the learning rate and number of epochs and selected the best checkpoint on the internal tuning set. We report the quadratic weighted kappa (QWK) for DR-CFP and AMD-CFP and macro-averaged F1 for AMD-OCT at both 224 px and 512/518 px resolutions. In Task 2, we used mean average precision (mAP) as the metric and evaluated at the same two resolutions (224 px and 512/518 px). Unless otherwise stated,

all mAP values are reported as the macro average across clinical sign categories within each dataset, and results for each individual sign are provided in the Supplementary.

Task 1: Point-supervised disease classification

DINOv3-Large and PE-Large demonstrated strong performance in internal validation, with QWK ranging from 0.866 to 0.917 in DR-CFP and from 0.857 to 0.909 in AMD-CFP, and F1 ranging from 0.885 to 0.916 in AMD-OCT. In the external validation, both models showed comparable performances in the OCT modality, with F1 ranging from 0.804 to 0.886 but declined in DR-CFP and AMD-CFP, where QWK ranged from 0.496 to 0.756. Among the ophthalmic foundation models, only the VisionFM variants and RETFound-DINOv2 maintained or improved their ranks on external validation. On external validation, RETFound-DINOv2 and VisionFM-CFP achieved QWK ranging from 0.548 to 0.689 in DR-CFP and from 0.855 to 0.919 in AMD-CFP. Both of these ranked in the top 3 at 224 and 512/518 px, except for VisionFM-CFP in DR-CFP at 512/518 px, which ranked 4th. On OCT, VisionFM-OCT ranked first at both resolutions, with F1 values of 0.842 and 0.904, respectively.

The remaining ophthalmic models showed variable or limited performance. For example, UrFound ranked in the upper half at 224 px on external validation, with a QWK of 0.541 in DR-CFP and 0.835 in AMD-CFP. However, its rank declined at 512/518 px. RETFound-Green mostly ranked in the lower half on DR-CFP and AMD-CFP, except in DR-CFP internal validation at 512/518 px. MIRAGE achieved only modest performance, even at its native resolution (512 px), with an F1 of 0.820 on external OCT. The performance declined further at 224 px, as expected, given its pretraining on 512 px inputs with 32×32 patches.

Figure 2 presents a summary of the classification performance on internal and external validation, measured as QWK for DR and AMD on CFP and as macro-averaged F1 for AMD on OCT. Detailed results for each individual dataset, including 95% confidence intervals, are shown in Supplementary Figs. 2–4.

Task 2: point-guided clinical sign-based image retrieval

The ophthalmic foundation models showed a clear and consistent advantage in clinical sign-based image retrieval. In within-dataset retrieval, VisionFM-CFP and VisionFM-OCT occupied the top ranking in nearly all DR-CFP, AMD-CFP, and AMD-OCT settings. Notably, VisionFM-CFP achieved mAP ranging from 0.395 to 0.443 for DR-CFP and from 0.369 to 0.390 for AMD-CFP, and VisionFM-OCT achieved mAP ranging from 0.462 to 0.472 for AMD-OCT, across resolutions. RETFound variants, including DINOv2 and Green for CFP and MAE for OCT, demonstrated meaningful performance across variants and resolutions, with mAP ranging from 0.289 to 0.410 for DR-CFP, from 0.270 to 0.374 for AMD-CFP, and from 0.269 to 0.280 for AMD-OCT. By comparison, DINOv3-Large and PE-Large, which had performed strongly on internal validation in Task 1, mostly ranked in the lower half in retrieval, with mAP ranging from 0.233 to 0.285 for DR-CFP, from 0.238 to 0.276 for AMD-CFP, and from 0.143 to 0.221 for AMD-OCT.

In cross-dataset retrieval, where each dataset had different characteristics, such as ethnicity and imaging device, the VisionFM variants remained dominant, achieving mAP ranging from 0.217 to 0.241 for DR-CFP, from 0.252 to 0.254 for AMD-CFP, and from 0.423 to 0.447 for AMD-OCT. The RETFound variants still showed strong performance, with mAP ranging from 0.104 to 0.218 for DR-CFP, from 0.155 to 0.226 for AMD-CFP, and from 0.300 to 0.310 for AMD-OCT. General-purpose vision foundation models remained well behind the VisionFM and RETFound variants in CFP retrieval, reaching only 0.086 to 0.104 mAP for DR-CFP and 0.114 to 0.152 mAP for AMD-CFP. In OCT, they reached 0.227 to 0.285 mAP.

The advantage of these ophthalmic models was most significant in a severely-skewed setting: cross-dataset retrieval on the UK Biobank population-screening cohort, where disease-related sign prevalence was low and highly imbalanced. Although the overall performance of every model decreased substantially, VisionFM-CFP and RETFound-DINOv2 still achieved notable mAP ranging from 0.026 to 0.073 for DR-CFP and from 0.090 to 0.113 for AMD-CFP across models and resolutions. By contrast, general-purpose foundation models, including DINOv3, reached mAP ranging from 0.002 to 0.006 for DR-CFP and from 0.049 to 0.068 for AMD-CFP. The overall performance is shown in Fig. 3, and results for each dataset and clinical sign are provided in Supplementary Figs. 5–7.

Three additional experiments were conducted. The first two analyses, presented in Supplementary Fig. 8, examined whether the ranking of foundation models is altered by modest adaptation via attention probing, or by using intermediate-layer rather than final-layer representations. Under attention probing, the gap separating the VisionFM variants and RETFound-DINOv2 from the other models remained substantial for both AMD-OCT resolutions and for DR-CFP at 224 resolution, but narrowed elsewhere. Nevertheless, the VisionFM variants and RETFound-DINOv2 remained among the top-performing models. Using intermediate-layer features improved performance for DINOv3 and PE-Large in particular, yet it remained below that of the top-performing models such as the VisionFM variants and RETFound-DINOv2. The third, presented in Supplementary Fig. 9, assessed whether the patch of maximum similarity actually contained the annotated clinical sign. Accuracy varied widely across models and tasks, ranging from 0.056 to 0.854. Notably, the VisionFM variants performed most consistently, with accuracy ranging from 0.317 to 0.433 in DR CFP, 0.583 to 0.807 in AMD CFP, and 0.658 to 0.737 in AMD OCT.

Effect of point annotation on classification and retrieval

In Task 1, point-supervision was associated with a statistically significant overall gain in classification performance in all six settings (all Benjamini-Hochberg adjusted $p < 0.001$), with overall effect sizes (Cohen's d_z) ranging from 0.296 to 0.874 across tasks and internal/external validation, though significance was less consistent at the level of individual backbones. The general-purpose foundation models showed larger deltas (with point supervision – without point supervision[naive]) than the ophthalmic foundation models. On the CFP tasks, their QWK deltas had a median of 0.021 (IQR 0.000 to 0.057) on DR-CFP and 0.026 (IQR 0.002 to 0.073) on AMD-CFP, while the AMD-OCT F1 delta had a median of 0.114 (IQR 0.051 to 0.201), the largest gain observed in Task 1.

Ophthalmic foundation models improved more modestly, with median QWK deltas of 0.017 (IQR – 0.005 to 0.056) on DR-CFP and 0.016 (IQR – 0.002 to 0.047) on AMD-CFP and a median AMD-OCT F1 delta of 0.033 (IQR 0.003 to 0.067).

The gain associated with point supervision was generally larger under external than internal validation. In ophthalmic foundation models, the median delta rose from 0.009 (IQR – 0.010 to 0.023) internally to 0.045 (IQR 0.006 to 0.087) externally on DR-CFP and from 0.012 (IQR – 0.003 to 0.026) to 0.025 (IQR 0.002 to 0.063) on AMD-CFP, whereas on AMD-OCT the two were comparable (0.033 [IQR 0.005 to 0.065] internally versus 0.030 [IQR 0.002 to 0.070] externally). General-purpose foundation models showed the same direction on DR-CFP (0.011 [IQR – 0.003 to 0.042] to 0.032 [IQR 0.008 to 0.065]) and AMD-OCT (0.100 [IQR 0.047 to 0.196] to 0.128 [IQR 0.068 to 0.206]), while on AMD-CFP the medians were similar and the external distribution was wider (0.022 [IQR 0.005 to 0.068] internally versus 0.030 [IQR – 0.007 to 0.074] externally).

In Task 2, point annotation was associated with a statistically significant overall gain in retrieval performance in all six settings (DR-CFP, AMD-CFP, and AMD-OCT, each evaluated within- and cross-datasets; all Benjamini-Hochberg adjusted $p < 0.001$), with overall effect sizes (Cohen's d_z) ranging from 0.523 to 1.288, though the size of this gain varied substantially across foundation models and individual clinical signs. Ophthalmic foundation models benefited more than general-purpose models in all three tasks (median mAP delta 0.021 [IQR 0.000 to 0.086] versus 0.003 [IQR 0.000 to 0.033] on DR-CFP, 0.068 [IQR 0.016 to 0.130] versus 0.047 [IQR 0.014 to 0.099] on AMD-CFP, and 0.038 [IQR 0.000 to 0.173] versus 0.023 [IQR 0.002 to 0.062] on AMD-OCT), although the benefit was concentrated in a few models. VisionFM variants showed the largest delta across all three tasks (0.079 [IQR 0.027 to 0.169] on DR-CFP, 0.119 [IQR 0.069 to 0.170] on AMD-CFP, and 0.174 [IQR 0.058 to 0.257] on AMD-OCT), followed by RETFound-DINOv2 on DR-CFP and AMD-CFP (0.073 [IQR 0.015 to 0.179] and 0.098 [IQR 0.054 to 0.198], respectively).

The delta also varied by category of clinical sign. On DR-CFP the largest gains were for hard exudates (median 0.135 [IQR 0.079 to 0.214] within-datasets and 0.031 [IQR 0.006 to 0.150] cross-datasets) and IRMA (0.081 [IQR 0.054 to 0.162] and 0.013 [IQR 0.000 to 0.079]), the same signs that accounted for most of the advantage of the leading models, whereas neovascularisation improved only marginally cross-datasets (0.001 [IQR 0.000 to 0.009]) and venous beading showed a small but significant reduction within-datasets (-0.018 [IQR -0.029 to -0.013], $d_z = 1.702$), plausibly because a point provides an ambiguous signal for a subtle change affecting only a short segment of an otherwise continuous vessel. On AMD-CFP every sign improved significantly in both settings, led by hard exudates and fibrosis. On AMD-OCT the benefit was largely confined to focal fluid and material: IRF showed the largest delta (0.128 [IQR 0.072 to 0.265] within-datasets and 0.117 [IQR 0.041 to 0.232] cross-datasets), SRF and SHRM improved significantly but modestly in both settings, and PED, drusen, and retinal pigment epithelium/ellipsoid zone change did not reach significance in either setting. Further details on Tasks 1 and 2 are provided in Figs. 4–6, respectively.

Discussion

In this study, we introduced a point-annotation-based framework for fine-grained evaluation of ophthalmic foundation models beyond conventional classification. General-purpose vision foundation models showed adaptability comparable to that of ophthalmic models in classification, whereas the benefit of ophthalmic pretraining emerged clearly when feature representations were directly probed for their ability to retrieve clinical signs. The results also indicate that point annotation provides meaningful supervision during training and is important for clinical sign retrieval, rather than relying on image-only retrieval.

General-purpose vision foundation models performed well across a wide range of general-domain tasks, and their demonstrated versatility raises questions about the need for domain-specific ophthalmic foundation models. Previous evaluations in the medical field, including ophthalmology, have reported that general-purpose foundation models can achieve comparable performance,²⁴ and recent head-to-head comparisons in ophthalmology have reached similar conclusions.^{25,26} Our Task 1 results are consistent with these reports; across all fundus and OCT tasks, general-purpose foundation models showed adaptability comparable to or exceeding that of ophthalmic foundation models under supervised fine-tuning. The competitiveness of general-purpose models was particularly significant in the OCT modality, whereas DINOv3 and PE-Large ranked among the top 3 in both internal and external validation. These two models also performed strongly on the DR-CFP and AMD-CFP tasks under internal validation. However, this picture changed under external validation. RETFound-DINOv2 and the VisionFM variants, both pre-trained on large-scale ophthalmic imaging datasets, improved their rankings. This suggests that ophthalmic pretraining can confer a generalisation advantage on unseen data. Such robustness likely reflects pretraining on large-scale ophthalmic images spanning diverse camera types and ethnic populations, which are factors known to affect performance on unseen data.²⁷

Ophthalmic foundation models showed a clear advantage in Task 2 (point-guided clinical sign retrieval), which directly probed sign-level feature encoding. The VisionFM variants ranked first in both within- and cross-dataset settings for DR-CFP, AMD-CFP, and AMD-OCT, with a single exception of within-dataset DR-CFP at 224 px, where RETFound-DINOv2 ranked first. This consistency suggests that VisionFM relies less on dataset-specific context.²⁸

By contrast, several ophthalmic and general-purpose foundation models showed limited retrieval performance in Task 2 despite performing competitively in Task 1. This divergence reflects the distinct properties that the two tasks probe: adaptability under supervised fine-tuning in Task 1 and the intrinsic informativeness of the pretrained representations in Task 2. PE-Large, for example, performed well in Task 1 classification but fell to near the bottom in Task 2 retrieval across all modalities. Such a discrepancy indicates that strong classification performance can be achieved through fine-tuning even when the pretrained features fail at retrieval, one of core aspects of a foundation model's representational capability.²⁹ The gap was most pronounced for the general-purpose models, where strong adaptability under fine-tuning did not extend to the pretrained representations, which transferred only partially to ophthalmic imaging.

The capability assessed in Task 2 is directly relevant to clinical decision support: retrieving prior cases that share the same specific clinical sign, such as a small retinal haemorrhage in early DR or subretinal fluid in AMD, provides clinicians with case-based evidence for anticipating the likely clinical course,

planning treatment, and counselling patients.³⁰ Beyond this, the task may be relevant to reliability in downstream tasks that are based on clinical signs, as it provides one of several measures of clinical sign understanding.

Existing benchmarks for ophthalmic foundation models have focused primarily on classification and segmentation. Moving beyond classification, RetiZero introduced disease-level image retrieval,¹⁰ and OCT-CUBE introduced same-eye retrieval across different modalities (OCT and IR) as evaluation tasks.⁹ However, they operate at the level of disease categories rather than the specific clinical signs that underlie each diagnosis, and retrieval based on the whole image may be confounded by factors unrelated to the intended target, such as camera device or patient demographics. Retrieval at the sign level is more challenging; given a single query image, a model cannot infer which sign is intended. For example, a query image containing cotton wool spots may also exhibit haemorrhages or hard exudates, and a model without explicit sign guidance may retrieve images that match these co-occurring clinical signs rather than the intended target. Point annotations resolve this ambiguity by precisely indicating which sign is being queried; Task 2 uses this to directly test whether a model's pretrained features encode individual clinical signs, such that a query image retrieves other images with the same sign.

Our experiments also demonstrated the value of point annotations. Notably, in Task 1, the advantage of point-supervision may indicate that it helps the model learn clinically meaningful signals rather than shortcut cues, such as device-specific appearance. For example, in the OCT modality, point-supervised models showed improved performance on the external dataset of Heidelberg Spectralis images despite having been trained on Topcon Triton images, suggesting that point-supervision could aid cross-device generalisation. This is consistent with earlier findings in dermatology, where supervising attention maps with segmentation masks reduced dataset-specific bias.^{32, 33} In Task 2, point annotations provided a different benefit: they resolved the inherent ambiguity of sign-level retrieval by specifying which sign is intended, enabling gains that a query image alone cannot achieve.

Conceptually, our study provides a rationale for an established line of research into the evaluation of foundation models. Earlier studies^{16, 34} made clinical signs and features explicit by segmenting or detecting them before diagnosis so that the model's decision relied on those findings rather than on shortcut cues. Our framework serves the same purpose as point annotation, indicating to the model which sign is of interest. Our results suggest a potential role for point annotation as a lightweight alternative to dense segmentation labels,³⁵ which can improve both classification generalisation and sign-level retrieval, particularly in data-scarce or rare-disease scenarios. More broadly, these findings argue that sign-level capability should be assessed directly when foundation models are evaluated and that pretraining objectives could be designed to explicitly encourage such features.

Several limitations should be considered when interpreting these findings. First, the evaluated models differed simultaneously in model capacity, pretraining data, and pretraining strategy, making it difficult to disentangle their individual contributions. For example, most ophthalmic foundation models were trained at 224×224 px, whereas DINOv3-Large and EUPE-Base were trained to 512×512 px, and MIRAGE uses a 32×32 patch that may degrade performance when evaluated at 224×224 px. Furthermore, rotary positional embeddings^{1,4,5} (RoPEs), which have been adopted by the latest general-purpose foundation models and are more robust to changes in input resolution than conventional positional embeddings, are not used in any of the ophthalmic foundation models evaluated here. Establishing causal relationships would require training foundation models under controlled conditions, varying each factor independently while keeping the others fixed. Second, vision transformers process images in fixed-size patches (typically 14×14 or 16×16 px),³⁶ which limits the spatial resolution of the learned representation. Small or closely spaced lesions that fall within a single patch are registered as one, which may dilute the fine-grained positional information regarding clinical signs and understate their true density. Third, our evaluation focused on two diseases (DR and AMD) across two imaging modalities (fundus photography and OCT). Although we included multiple external validation datasets spanning different populations and devices, further work is required to determine whether these findings generalise to other ophthalmic diseases and imaging modalities. Finally, although point annotation is far less costly than pixel-level segmentation masks, it still requires expert graders and remains labour-intensive, which may limit the scalability of this evaluation framework to substantially larger benchmarks.

In summary, this study introduced a point annotation-based framework that moves ophthalmic foundation model evaluation beyond disease-level classification toward fine-grained recognition of individual clinical signs. Our findings show that strong classification performance does not guarantee sign-level feature encoding and that conventional benchmarks can therefore obscure fundamental limitations in what a model has actually learned. As foundation models are increasingly positioned as universal backbones for both general and ophthalmic applications, evaluation should go beyond classification and segmentation accuracy to determine whether a model's learned image representations are grounded in the clinical observations that underlie diagnosis.

Methods

Task definitions

We aimed to evaluate DR severity classification for CFP based on the international clinical diabetic retinopathy (ICDR) severity scale¹² with five classes and AMD classification for CFP based on the Beckman clinical classification of AMD with four classes.³⁷ For OCT, we aimed to evaluate a three-class AMD classification: normal, dry/inactive AMD, or active/exudative AMD.¹⁵ Although the DR-CFP and AMD-CFP tasks involve ordinal disease severity, the AMD-OCT task does not, as reliable staging into early, intermediate, and late AMD was not possible from central B-scans alone. Therefore, we used broader clinical categories of normal, dry or inactive AMD (including neovascular AMD without signs of exudative activity), and active neovascular AMD.

Overview of datasets

We assembled 10 datasets for this study. Four datasets (HYU-DR-CFP, HYU-AMD-CFP, HYU-AMD-OCT, and Yonsei-AMD-OCT; all from South Korea) were curated in-house from two tertiary hospitals (Hanyang University Guri Hospital and Yonsei University Severance Hospital). The training split of the ADAM

dataset (China)³⁸ and full IDRiD dataset (India)³⁹ were used with additional annotation. Four population-based datasets (UK Biobank [UK], Beijing Eye Study [BES, China], Ural Eye and Medical Study [UEMS, Russia], and Central India Eye and Medical Study [CIEMS, India])⁴⁰ were also used.

The three HYU datasets were collected from the ophthalmology outpatient clinic of Hanyang University Guri Hospital, a tertiary medical centre, between March 1, 2020, and March 1, 2024. The Yonsei-AMD-OCT dataset was collected from the ophthalmology outpatient clinic of Yonsei University Severance Hospital, a tertiary medical centre, between March 1, 2022, and March 1, 2023.

For the HYU-DR-CFP dataset, CFP images were restricted to patients with available glycated haemoglobin (HbA1c) measurements; to follow the intended use of DR screening tools, analyses were further restricted to participants with diabetes mellitus ($\text{HbA1c} \geq 6.5\%$). Likewise, the other datasets used in Task 1 (DR-CFP) were filtered to include only patients with diabetes mellitus ($\text{HbA1c} \geq 6.5\%$). For the HYU-AMD-CFP and HYU-AMD-OCT datasets, we included only patients who were treatment-naïve at the time of image acquisition to ensure that the fundus photographs and OCT B-scans reflected the natural, untreated appearance of each AMD stage. During image collection, fundus photographs and OCT B-scans of insufficient quality for reliable grading due to severe media opacity (e.g., cataract or vitreous opacities), off-centre acquisition, or blink, motion, and mirror artifact were dropped. Patients were also excluded if they had a history of other ocular diseases that could confound DR or AMD grading, including high myopia, glaucoma, retinal vascular occlusion, or active uveitis. The same criteria were applied to the Yonsei-AMD-OCT dataset.

IDRiD, in which all patients have diabetes, was used after filtering out low-quality CFPs, with clinical signs additionally annotated on top of the existing class labels. For ADAM, the training split (of the three original splits) was re-annotated into four classes, again with clinical signs additionally annotated. Detailed dataset information, including fixation protocols, CFP and OCT devices, number of patients and images used in this study, and demographic summaries, is summarised in Table 1.

Table 1
Summary of the datasets used for Task 1 (disease classification)

Dataset	Country	Setting	Age (y)	Sex (% Male)	Imaging device	Patients (n)	Images (n)	Split	Class distribution (n)			
DR-CFP									No DR	Mild NPDR	Moderate NPDR	Severe NPDR
HYU-DR-CFP (internal)	South Korea	Clinic	54.5 ± 14.4	46.3%	DRI OCT Triton (Topcon, Tokyo, Japan)	112	416	Training	169	100	87	34
						34	124	Tuning	52	37	15	6
						80	293	Validation	119	70	53	30
						226	833	Total	340	207	155	70
IDRiD (external)	India	Clinic	NA	NA	VX-10 (Kowa, Nagoya, Japan)	NA	490	External validation	165	25	160	90
CIEMS (external)	India	Population	52.7 ± 11.8	47.3%	FF450 (Carl Zeiss Meditec, Jena, Germany)	146	461	External validation	430	10	4	3
BES (external)	China	Population	62.1 ± 9.8	43.9%	CR6-45NM (Canon, Tokyo, Japan)	494	1766	External validation	1492	132	82	28
UEMS (external)	Russia	Population	62.2 ± 8.4	26.1%	Visucam 500 (Carl Zeiss Meditec, Jena, Germany)	282	620	External validation	510	73	33	1
AMD-CFP									No AMD	Early AMD	Intermediate AMD	Late AM
HYU-AMD-CFP (internal)	South Korea	Clinic	63.2 ± 14.6	48.4%	DRI OCT Triton (Topcon, Tokyo, Japan)	189	316	Training	126	36	49	105
						57	94	Tuning	37	10	15	32
						133	219	Validation	99	23	41	56
						379	629	Total	262	69	105	193
ADAM (external)	China	Clinic	NA	NA	Visucam 500 (Carl Zeiss Meditec, Jena, Germany), CR-2 (Canon, Tokyo, Japan)	NA	345	External validation	281	20	7	37
AMD-OCT									No AMD	AMD without signs of active exudation	AMD with signs of active exudation	

Datasets are grouped by task: DR grading on colour fundus photographs (DR-CFP), AMD grading on CFP (AMD-CFP) and AMD classification on OCT (AMD-OCT). For each task the internal dataset (HYU) is listed first, broken down into its training, tuning and validation splits with the corresponding total, followed by the external validation datasets. Internal splits were partitioned at the patient level, so no patient contributes images to more than one split. Age and sex were derived at the patient level; age is reported as mean ± SD and sex as the percentage of male patients. Class distribution is the number of images per class, as the number of classes differs by task: five for DR-CFP, four for AMD-CFP and three for AMD-OCT. Model performance was evaluated using quadratic weighted kappa for DR-CFP and AMD-CFP and macro-averaged F1 for AMD-OCT. NA indicates that information is not available.

Abbreviations: ADAM, Automatic Detection challenge on Age-related Macular degeneration; AMD, age-related macular degeneration; BES, Beijing Eye Study; colour fundus photograph; CIEMS, Central India Eye and Medical Study; DR, diabetic retinopathy; IDRiD, Indian Diabetic Retinopathy Image Dataset; NPDR, non-proliferative diabetic retinopathy; OCT, optical coherence tomography; PDR, proliferative diabetic retinopathy; SD, standard deviation; UEMS, Ural Eye and Medical Study.

Dataset	Country	Setting	Age (y)	Sex (% Male)	Imaging device	Patients (n)	Images (n)	Split	Class distribution (n)		
HYU-AMD-OCT (internal)	South Korea	Clinic	64.6 ± 9.8	47.5%	DRI OCT Triton (Topcon, Tokyo, Japan)	308	485	Training	320	32	133
						93	146	Tuning	96	11	39
						216	348	Validation	232	27	89
						617	979	Total	648	70	261
Yonsei (external)	South Korea	Clinic	74.7 ± 12.1	49.3%	SPECTRALIS OCT (Heidelberg Engineering, Heidelberg, Germany)	339	483	External validation	155	188	140

Datasets are grouped by task: DR grading on colour fundus photographs (DR-CFP), AMD grading on CFP (AMD-CFP) and AMD classification on OCT (AMD-C). For each task the internal dataset (HYU) is listed first, broken down into its training, tuning and validation splits with the corresponding total, followed by the external validation datasets. Internal splits were partitioned at the patient level, so no patient contributes images to more than one split. Age and sex were derived at the patient level; age is reported as mean ± SD and sex as the percentage of male patients. Class distribution is the number of images per class, as the number of classes differs by task: five for DR-CFP, four for AMD-CFP and three for AMD-OCT. Model performance was evaluated using quadratic weighted kappa for DR-CFP and AMD-CFP and macro-averaged F1 for AMD-OCT. NA indicates that information is not available.

Abbreviations: ADAM, Automatic Detection challenge on Age-related Macular degeneration; AMD, age-related macular degeneration; BES, Beijing Eye Study; colour fundus photograph; CIEMS, Central India Eye and Medical Study; DR, diabetic retinopathy; IDRI, Indian Diabetic Retinopathy Image Dataset; NPDR, non-proliferative diabetic retinopathy; OCT, optical coherence tomography; PDR, proliferative diabetic retinopathy; SD, standard deviation; UEMS, Ural Eye and Me Study.

Image grading and annotation protocol

Image grading and point annotation for clinical signs were performed separately for three tasks: DR grading in CFPs, AMD grading in CFPs, and AMD grading in OCT central B-scans. OCT evaluation was restricted to central B-scans to ensure comparability across models, particularly with established 2D retinal foundation models such as RETFound-MAE and VisionFM-OCT, which are designed and evaluated using single frames. Annotation was performed for the HYU-DR-CFP, HYU-AMD-CFP, HYU-AMD-OCT, Yonsei-AMD-OCT, ADAM, and IDRI datasets; for IDRI, only clinical signs were annotated, as the established classification labels were used.

All classification gradings were performed by one board-certified retina specialist and one reading centre-certified retina fellow. The two annotators graded each image for each class independently, and disagreements were adjudicated by an independent retina specialist with more than 15 years of clinical experience. Point annotation was then performed at a relatively coarse level compared with segmentation, in keeping with the patch-wise behaviour of vision transformers (typically 16×16 or 14×14 px). Each image was resized to 1024×1024 and overlaid with a grid of 16×16 px cells, and a point was placed within any cell in which the corresponding clinical sign was present. Across all three tasks, images were excluded when a diagnosis could not be established from the imaging modality alone or when additional examinations would be required to reach a diagnosis. For example, in the CFP task, subtle choroidal neovascularization (CNV) with active exudation may not produce a discernible surface correlate on CFP, precluding reliable classification as late AMD without OCT confirmation. Similarly, in the OCT task, the use of a single central B-scan meant that AMD-related changes could be absent from the selected frame or insufficiently represented in their extent and distribution, precluding a reliable AMD diagnosis in some cases.

For DR grading in CFPs, each image was assigned an overall severity grade according to the ICDR severity scale: no DR (normal), mild non-proliferative diabetic retinopathy (NPDR), moderate NPDR, severe NPDR, or proliferative DR (PDR). Seven clinical signs were then annotated point-wise for each image: haemorrhages, neovascularisation, cotton-wool spots, hard exudates, venous beading, fibrovascular membrane (fibrosis), and intraretinal microvascular abnormalities (IRMA).

For AMD grading in CFPs, each image was classified according to a modified Beckman clinical classification of AMD into one of four mutually exclusive classes: normal (no AMD-related pathology or only age-related changes; no lesion-level annotation required), early AMD (small/intermediate drusen), intermediate AMD (large drusen and/or pigmentary changes), or late AMD (advanced features including atrophy, haemorrhage, fibrosis, CNV/active exudation, and/or hard exudates). Eight clinical signs were then annotated point-wise: small/intermediate drusen, large drusen, pigmentary changes, atrophy, haemorrhage, subretinal fibrosis (fibrosis), CNV including active exudation, and hard exudates. Because exact drusen measurements were unavailable during grading, drusen size was estimated relative to the venous vessel diameter at the optic disc, with smaller drusen annotated as small/intermediate and larger drusen as large.³⁷

For AMD grading in OCT central B-scans, each image was classified as normal, dry/inactive AMD, or active/exudative AMD. Normal was defined as no AMD-related pathology or only age-related changes and required no lesion-level annotation. Dry/inactive AMD was defined as the presence of at least one structural biomarker not indicative of active exudation—drusen, pigment epithelial detachment (PED), subretinal fibrosis, or RPE/ellipsoid zone change (a composite category for outer retinal or RPE abnormalities, including atrophic changes, double-layer sign, and subtle RPE/ellipsoid zone alterations). This means that this category also included cases of neovascular AMD currently without signs of active exudation. Active/exudative AMD required at least one activity-related biomarker (subretinal fluid (SRF), intraretinal fluid (IRF), and/or subretinal hyperreflective material (SHRM)), consistent with the established

role of retinal fluid in OCT as a marker of neovascular AMD activity; SHRM was included as a characteristic OCT biomarker that may contain exudative, haemorrhagic, or MNV components. Seven clinical signs were then annotated point-wise: SRF, IRF, SHRM, drusen, PED, subretinal fibrosis, and RPE/ellipsoid zone changes.

BES, CIEMS, and UEMS were each labelled independently by board-certified ophthalmologists with more than 20 years of experience. However, as these datasets were graded for different purposes according to their respective study protocols, not all clinical signs and disease labels were available. The available tasks, clinical signs, and modalities for each dataset are listed in Table 2.

Table 2
Summary of the datasets used for Task 2 (clinical sign retrieval)

Dataset	Country	Setting	Age (y)	Sex (% Male)	Imaging device	Patients (n)	Images (n)	Annotated images per clinical sign (n)				
DR-CFP								Haemorrhage	Hard exudates	Cotton wool spots	IRMA	Vertical lines
HYU-DR-CFP (internal)	South Korea	Clinic	61.1 ± 10.3	60.2%	DRI OCT Triton (Topcon, Tokyo, Japan)	226	833	488	193	206	58	38
IDRiD (external)	India	Clinic	NA	NA	VX-10 (Kowa, Nagoya, Japan)	NA	490	330	230	156	79	90
CIEMS (external)	India	Population	49.4 ± 13.0	47.0%	FF450 (Carl Zeiss Meditec, Jena, Germany)	3,350	12,419	NA	13	NA	6	NA
BES (external)	China	Population	64.3 ± 9.7	43.7%	CR6-45NM (Canon, Tokyo, Japan)	3,402	13,222	NA	NA	153	NA	NA
UK Biobank (external)	United Kingdom	Population	55.6 ± 8.3	44.0%	3D OCT-1000 Mark II (Topcon, Tokyo, Japan)	38,634	76,696	248	122	80	15	8
AMD-CFP								Small/intermediate drusen	Large drusen	Pigmentary changes	Atrophy	CN
HYU-AMD-CFP (internal)	South Korea	Clinic	71.0 ± 9.1	59.6%	DRI OCT Triton (Topcon, Tokyo, Japan)	379	629	257	135	64	32	37
ADAM (external)	China	Clinic	NA	NA	Visucam 500 (Carl Zeiss Meditec, Jena, Germany), CR-2 (Canon, Tokyo, Japan)	NA	345	40	10	13	22	6
BES (external)	China	Population	64.3 ± 9.7	43.7%	CR6-45NM (Canon, Tokyo, Japan)	3,402	13,222	3,556	2,230	584	21	NA
UK Biobank (external)	United Kingdom	Population	55.6 ± 8.3	44.0%	3D OCT-1000 Mark II (Topcon, Tokyo, Japan)	38,634	76,696	7,264	1,316	NA	14	NA
AMD-OCT								PED	SRF	IRF	SHRM	Subfoveal fluid
HYU-AMD-OCT (internal)	South Korea	Clinic	66.7 ± 11.4	53.6%	DRI OCT Triton (Topcon, Tokyo, Japan)	617	979	262	195	93	93	30

Datasets are grouped by task: DR on colour fundus photographs (DR-CFP), AMD on CFP (AMD-CFP) and AMD on OCT (AMD-OCT). For each task the internal datasets used for cross-dataset retrieval. Age is reported as mean ± SD and sex as the percentage of male patients. Counts for each clinical sign are the number present. The annotated clinical signs differ across datasets; a dash indicates that the sign was not annotated in that dataset and was therefore excluded from evaluation. The mean average precision (mAP), computed separately for each clinical sign and then averaged without weighting across signs and, where mAP is NA indicates that information is not available.

Abbreviations: AMD, age-related macular degeneration; CFP, colour fundus photograph; CNV, choroidal neovascularization; DR, diabetic retinopathy; IRF, intraretinal fluid; mAP, mean average precision; OCT, optical coherence tomography; PED, pigment epithelial detachment; RPE, retinal pigment epithelium; SHRM, subfoveal fluid.

Dataset	Country	Setting	Age (y)	Sex (% Male)	Imaging device	Patients (n)	Images (n)	Annotated images per clinical sign (n)				
Yonsei (external)	South Korea	Clinic	74.7 ± 12.1	49.3%	SPECTRALIS OCT (Heidelberg Engineering, Heidelberg, Germany)	339	483	234	84	75	5	65
Datasets are grouped by task: DR on colour fundus photographs (DR-CFP), AMD on CFP (AMD-CFP) and AMD on OCT (AMD-OCT). For each task the internal datasets used for cross-dataset retrieval. Age is reported as mean ± SD and sex as the percentage of male patients. Counts for each clinical sign are the number present. The annotated clinical signs differ across datasets; a dash indicates that the sign was not annotated in that dataset and was therefore excluded from evaluation using mean average precision (mAP), computed separately for each clinical sign and then averaged without weighting across signs and, where mAP is NA indicates that information is not available.												
Abbreviations: AMD, age-related macular degeneration; CFP, colour fundus photograph; CNV, choroidal neovascularization; DR, diabetic retinopathy; IRF, intra-retinal fluid; mAP, mean average precision; OCT, optical coherence tomography; PED, pigment epithelial detachment; RPE, retinal pigment epithelium; SHRM, subretinal fluid.												

For UK Biobank, we used labels from previous work, which were established by the Central Administrative Research Facility-certified graders and quality-assured by clinicians of the Network of Ophthalmic Reading Centres UK. As with the aforementioned datasets, because grading was performed for different purposes according to the study protocol, not all clinical signs and disease labels were available; the available tasks, clinical signs, and modalities are likewise listed in Table 2.

Institutional data collection and ethics

The collection and analysis of the HYU-DR-CFP, HYU-AMD-CFP, and HYU-AMD-OCT datasets were approved by the Institutional Review Board of Hanyang University Guri Hospital (IRB No. 2024-12-007-001) and conducted in accordance with the tenets of the Declaration of Helsinki. Similarly, the collection and analysis of the Yonsei-AMD-OCT dataset were approved by the Institutional Review Board of Yonsei University Severance Hospital (IRB No. 4-2025-1239). The requirement for informed consent was waived owing to the retrospective nature of the study, whereas informed consent was obtained from participants or their legal representatives in the UEMS, BES, and CIEMS cohorts.

Dataset preprocessing

For image quality control, we applied AutoMorph⁴¹ to the ADAM, IDRiD, UEMS, BES, and CIEMS datasets, and retained only images graded as "Good" or "Usable". UK Biobank was filtered using the gradability column provided in previous work.⁴² For HYU-DR-CFP, HYU-AMD-CFP, HYU-AMD-OCT, and Yonsei-AMD-OCT, low-quality images were excluded during data collection. On CFP, low-quality was defined as incomplete visibility of the optic disc or macula, or severe blur, dense media opacity, or marked over- or under-exposure precluding assessment of key lesions, and for OCT as low signal strength or excessive noise, blur, motion, blink or cropping artifacts and poor centration. In Task 1, the DR-CFP task included only participants with HbA1c ≥ 6.5%, in line with the diagnostic criterion for diabetes mellitus. The HYU datasets used as the internal validation set in Task 1 were split into training, tuning, and validation subsets in a 50:15:35 ratio, stratified by target class, and partitioned at the patient level to prevent leakage. All images were pre-processed to 1024×1024 and then transformed to the input resolution specified during training.

Foundation models

We evaluated 11 publicly available foundation models, including three generic, five CFP-based, and three OCT-based. OCT-CUBE was excluded because it is based on full 3D OCT volume scans.

Among the general-purpose vision foundation models, DINOv3¹ demonstrated strong performance across various tasks, and it was trained on 1.689 billion images curated from an initial pool of 17 billion images. Perception Encoder⁴ was designed for image and video understanding and trained with a contrastive vision–language objective on 5.4 billion image–text pairs and 22 million videos paired with text. Efficient Universal Perception Encoder⁵, which is the most recent of the evaluated models, was developed to provide powerful and universal representations. It was trained by distilling knowledge from multiple teacher models, including Perception Encoder and DINOv3, using a multi-resolution distillation strategy at resolutions of 256 × 256, 384 × 384, and 512 × 512.

Among the CFP-based foundation models, RETFound-DINOv2⁷ was built on a ViT-Large backbone initialised from DINOv2 weights and further pretrained on 904,170 fundus photographs from Moorfields Eye Hospital at a resolution of 224 × 224. VisionFM-CFP;²³ based on a ViT-Base backbone, was pretrained using the image BERT pre-training with online tokenizer (iBOT) method on 1,010,293 fundus photographs. RETFound-Green²² employs a ViT-Small backbone initialised from DINOv2 pretrained weights and was trained at 392 × 392 resolution using the authors' proposed token-reconstruction strategy. UrFound⁸ uses a ViT-Base backbone and was pretrained with a knowledge-guided approach that combines masked image modelling and masked language modelling on retinal images and their paired textual annotations, using 103,786 fundus photographs and 83,484 OCT images. While OCT was part of its pretraining data, UrFound's original evaluation relied predominantly on fundus, with only one of the eight downstream datasets (OCTID) being OCT; we therefore grouped it with the CFP-based models. RetiZero¹⁰ was built upon RETFound-MAE (ViT-Large) and pretrained on 341,896 fundus photographs at 224 × 224 resolution, with the image encoder adapted using low-rank learnable factors and aligned with text.

Among the OCT-based foundation models, RETFound-MAE⁶ was based on a ViT-Large backbone pretrained with a masked autoencoder on 736,442 OCT slices from Moorfields Eye Hospital and the Kermany dataset at a resolution of 224 × 224. VisionFM-OCT²³, a ViT-Base model, was pretrained using the iBOT method on 1,420,357 OCT slices at 224 × 224 resolution. Notably, MIRAGE-Large¹¹ was trained under a multimodal scheme incorporating OCT, scanning

laser ophthalmoscopy (SLO), and OCT layer-segmentation masks derived from the OCT device. MIRAGE-Large uses a ViT-Large backbone with 512×512 inputs and a 32×32 patch size, whereas all other models employ 16×16 or 14×14 patches. Further details are provided in Supplementary Table 1.

Experimental Setup and Evaluation

In Task 1, we performed a hyperparameter sweep over learning rate and number of epochs and selected the best-performing checkpoint on the internal tuning set, using QWK as the primary metric for DR-CFP and AMD-CFP classification and macro F1 for AMD-OCT classification; the same metrics were used to report test performance. The search space comprised learning rates of $\{5e-5, 1e-4, 2e-4\}$ and training durations of $\{10, 30, 50\}$ epochs⁷, with AdamW⁴³ used as the optimiser in all runs. To ensure fair comparison, each model was assessed under two resolution settings: 224 and 512/518. Image augmentations, comprising random resized cropping (scale 0.9 to 1.0), horizontal and vertical flips, $\pm 20^\circ$ rotation, and brightness/contrast jitter, were applied during training, whereas only resizing was applied in the evaluation phase. For normalisation, PE-Large and RETFound-Green used mean 0.5 and std 0.5 for all channels, and MIRAGE used per-image min to max scaling; all other models used ImageNet mean/std. In Task 2, we used mAP as the primary metric. Each model was again assessed under the same two resolution settings: 224 and 512/518.

Evaluation framework: Task 1 (point-supervised disease classification)

Task 1 evaluated whether foundation models performed well in disease classification when trained with point-supervision. Each foundation model was used without architectural modification other than the addition of a linear classification head.^{36,44} From a single CFP image or OCT slice, the model outputs a predicted class label together with an attention map derived from the last attention block and averaged over all attention heads. For models with a class token, this map was taken as the attention weights between the class token and the image patch tokens. For models without a class token, the patch-to-patch attention map $\mathbb{R}^{HW \times HW}$ was averaged across all attending patches, yielding a per-patch distribution $\mathbb{R}^{HW}$ that indicates, on average, which patches the model relies on. The following two loss terms were used for fine-tuning:

$$L_{total} = L_{cls} + \lambda L_{point}$$

The first term, L_{cls} , is cross-entropy, which is the de facto standard for classification, as expressed by the following equation, where B is the batch size and C is the number of classes:

$$L_{cls} = -\frac{1}{B} \sum_{i=1}^B \sum_{c=1}^C y_{i,c} \log p_{i,c}$$

The second term, L_{point} , is a cosine-similarity loss adopted from prior work,⁴⁵ computed between the model attention map (a) and the Gaussian-smoothed point map (m), as expressed by the following equation:

$$L_{point} = \frac{1}{B} \sum_{i=1}^B (1 - \cos(a_i, m_i)) \cdot 1[c_i \neq 0]$$

This term encourages the model to attend to the regions identified by clinicians, with λ set to 1.0 following prior work. The loss is activated only for cases in which a clinical sign is present, and is therefore inactive for normal cases. The Gaussian-smoothed point map was generated by smoothing a binary point map, in which each pixel was set to one where any clinical sign had been annotated and to zero elsewhere, with a 7×7 Gaussian kernel. It was then average-pooled onto the $[H/P] \times [W/P]$ patch grid, matching the size of the attention map, and normalised so that the weights sum to one, where H and W denote the height and width of the input image in pixels, and P denotes the patch size. For the point-supervision ablation, λ was set to 0, so that the model was trained with cross-entropy alone.

Evaluation framework: Task 2 (point-guided clinical sign retrieval)

Unlike conventional classification and prognosis tasks, Task 2 evaluated whether clinical signs could be retrieved directly from the pretrained feature representations of the foundation models, without any task-specific adaptation. This form of feature representation evaluation is relevant to a range of downstream applications beyond retrieval, including few-shot learning in rare diseases, unsupervised discovery of disease subtypes through clustering, and multimodal systems, such as vision-language models, in which visual and textual representations must be aligned for the model to describe what it sees.

In this task, the query-set denotes the pool of images carrying point annotations, and these points specify the clinical sign to be retrieved. For each retrieval, a single image was drawn from the query-set as the query. The gallery-set denotes the pool of candidate images from which retrieval is performed. Features were taken from the frozen backbone as the spatial feature map immediately prior to any aggregation step, such as global average pooling, attention pooling, or class-token pooling. For an image of height H and width W with patch size P, this feature map is $F \in \mathbb{R}^{[H/P] \times [W/P] \times D}$, where D is the feature dimension.

For each query image, query patches were defined as follows. The binary point map of a given sign, set to one at an annotated point and zero elsewhere, was smoothed with a 7×7 Gaussian kernel, average-pooled onto the $[H/P] \times [W/P]$ patch grid, which matches the feature map after patchification, and normalised so that the weights sum to one. Patches whose weight w_i exceeded the uniform weight $1/([H/P] \times [W/P])$ were taken as the query patches, denoted by the index set I. Their feature vectors were $q_i = F_q(i) \in \mathbb{R}^D$ for $i \in I$, yielding the set of query patch features $Q = \{q_i\}_{i \in I}$.

Because the objective was to retrieve images containing the same clinical sign, not merely a similar overall appearance, similarity was computed at the patch level rather than the image level. For a query feature q_i and a gallery feature map $F_j \in \mathbb{R}^{[H/P] \times [W/P] \times D}$ extracted from gallery image j, the patch score s_j^i was defined as the maximum cosine similarity over all spatial positions of that feature map, as expressed by the following equation, where x and y index the spatial positions of the gallery image feature map:⁴⁶

$$s_j^i = \max_{x,y} \frac{q_i^T f_{j,x,y}}{\|q_i\| \|f_{j,x,y}\|}$$

The patch scores of a query image were then combined into a single image-level score as the weighted average of these patch-wise maxima, using the normalised weights w_i defined above, so that all points annotated for the sign of interest in that image, for example several hard exudates, were summarised into one score, with the weights reflecting both the proximity and the density of the annotated points, as expressed by the following equation:

$$S_j = \frac{\sum_{i \in I} w_i s_j^i}{\sum_{i \in I} w_i}$$

Gallery images were ranked in descending order of S . In the within-dataset setting, each HYU dataset (HYU-DR-CFP, HYU-AMD-CFP, and HYU-AMD-OCT) served as both the query-set and the gallery-set, and all images from the same patient as the query, including the query image itself, were excluded from the gallery-set. In the cross-dataset setting, each HYU dataset served as the query-set and the remaining datasets, such as UK Biobank, as the gallery-set; no images were excluded, because no patient overlap existed between the query-set and the gallery-set. Confidence intervals were estimated from 2,000 bootstrap resamples of the query patients, and performance was reported as mAP. A detailed illustration and description are provided in Supplementary Fig. 1.

To examine whether the limited sign-level performance of the general-purpose models in Task 2 reflected information that was either not encoded or not readily recoverable under direct cosine similarity, each model was additionally evaluated under attention probing, with each sign formulated as an independent binary classification task. For each foundation model and each sign, the frozen feature map was passed to a separate attention-pooling head that aggregated patch features into an image-level representation, which was then mapped to a single logit. The heads were implemented using the timm library, and the backbone remained frozen throughout, so that only the pooling heads were trained. Each head was trained for 50 epochs with binary cross-entropy loss, a batch size of 32, and AdamW (learning rate 2×10^{-4} , weight decay 1×10^{-2} , cosine learning-rate schedule), using patient-level stratified splits of 80% for training and 20% for tuning. Because the internal data were used entirely for training and tuning, the results of attention probing are reported for the external validation datasets only. We also repeated the Task 2 retrieval using representations taken from intermediate layers at 25%, 50%, and 75% of the network depth, keeping all other settings unchanged, to test whether the ranking of the models depended on the depth at which features were extracted. The results of both analyses are shown in Supplementary Fig. 8.

Additionally, we assessed pointing accuracy, that is, whether the retrieved patch corresponded to the queried clinical sign. For each query image, a patch-level similarity map was computed between every query patch and all patches of a given gallery image, and these maps were weighted averaged across the query patches to yield a single similarity map per gallery image. The patch with the highest averaged similarity was then selected, and pointing accuracy was defined as the proportion of selected patches that contained the queried sign. It was computed only on correct top-1 retrievals, that is, gallery images sharing the same clinical sign as the query, so that the metric reflected localisation quality rather than retrieval correctness. The evaluation was performed on the datasets containing point annotations, namely HYU-DR-CFP, IDRiD, HYU-AMD-CFP, ADAM, HYU-AMD-OCT, and Yonsei-AMD-OCT. Values were macro-averaged across clinical signs within each task and are reported in Supplementary Fig. 9

Statistical analysis

All metrics are reported with 95% confidence intervals (CIs) estimated by bootstrap resampling with 2,000 iterations. Resampling was performed at the patient level for Task 1 and at the patient level among query images for Task 2; for the two datasets without patient identifiers (IDRiD and ADAM), resampling was performed at the image level. CIs were taken as the 2.5th and 97.5th percentiles of the bootstrap distribution. In Task 1, paired differences in QWK and F1 were computed between the point-supervised and naive (no point supervision) conditions. Each pair was matched on the hyperparameter configuration, so that the two conditions differed only in the presence of point supervision. In Task 2, paired differences in mAP were computed between the point-guided and naive (no point-guidance) conditions. The null hypothesis of zero median difference was tested two-sided using the Wilcoxon signed-rank test, with p values adjusted by the Benjamini–Hochberg procedure ($q < 0.05$ considered significant). Effect sizes were quantified as Cohen's d_z for paired samples, defined as the mean paired difference divided by its standard deviation, and are reported alongside the corresponding p values. Analyses were performed in Python 3 using SciPy.

Data and code availability

The publicly available datasets used in this study can be accessed as follows: ADAM (<https://amd.grand-challenge.org/>; also available via IEEE DataPort at <https://doi.org/10.21227/dt4f-rt59>) and IDRiD (<https://idrid.grand-challenge.org/>). The UEMS, CIEMS, and BES datasets are not publicly available due to participant privacy and institutional data-sharing agreements, but de-identified data may be made available from the corresponding investigators (M.M.B./G.M.K., V.N., and J.B.J./Y.Wa., respectively) upon reasonable request and approval by the relevant institutional review boards. The evaluation framework and datasets curated in this study, comprising the Hanyang University Guri Hospital datasets (HYU-DR-CFP, HYU-AMD-CFP, and HYU-AMD-OCT) and the Yonsei-AMD-OCT dataset, are available at https://github.com/hyunminkim0625/PointGuided_Benchmark under a CC BY-NC 4.0 license.

The publicly available foundation models evaluated in this study can be obtained from their respective repositories: RETFound-DINOv2 and RETFound-MAE (<https://github.com/rmaphoh/RETFound>), VisionFM-CFP and VisionFM-OCT (<https://github.com/ABILab-CUHK/VisionFM>), RETFound-Green (https://github.com/justinengelmann/RETFound_Green), RetiZero (<https://github.com/Looking9218/RetiZero>), UrFound (<https://github.com/yukkai/UrFound>), MIRAGE (<https://github.com/j-morano/MIRAGE>), DINOv3 (<https://github.com/facebookresearch/dinov3>), Perception Encoder (https://github.com/facebookresearch/perception_models), and Efficient Universal Perception Encoder

(<https://github.com/facebookresearch/EUPE>). Image quality control was performed using the automated retinal image analysis tool AutoMorph (<https://github.com/rmaphoh/AutoMorph>).

Abbreviations

ADAM
Automatic Detection challenge on Age-related Macular degeneration
AMD
age-related macular degeneration
BES
Beijing Eye Study
CFP
colour fundus photograph
CIEMS
Central India Eye and Medical Study
DR
diabetic retinopathy
HYU
Hanyang University
ICDR
International Clinical Diabetic Retinopathy
IDRiD
Indian Diabetic Retinopathy Image Dataset
NPDR
non-proliferative diabetic retinopathy
OCT
optical coherence tomography
PDR
proliferative diabetic retinopathy
UEMS
Ural Eye and Medical Study.

Declarations

Funding

This work was supported by the Engineering and Physical Sciences Research Council [grant number UKRI1662].

Competing Interests

P.A.K. is a co-founder of Cascader Ltd. and has acted as a consultant for Skleo Health, insitro, Retina Consultants of America, Roche, Boehringer-Ingelheim, and Bitfount, and is an equity owner in Big Picture Medical. He has received speaker fees from Zeiss, Théa, Apellis, and Roche, travel support from Bayer and Roche, and grant funding from Roche. He has attended advisory boards for Topcon, Bayer, Boehringer-Ingelheim, and Roche. J.B.J. holds European patents EP 3 271 392 and WO 2025/056713 A1, Japanese patent JP 2021-119187, and U.S. Patent No. US 12,024,557 B2. J.H. has received travel support from Bayer and Roche and speaker fees from Zeiss, Roche, and Bayer. S.P.J. holds European patent EP 3 271 392, Japanese patent JP 2021-119187, and U.S. Patent No. US 12,024,557 B2: Agents for use in the therapeutic or prophylactic treatment of myopia or hyperopia; and European patent application WO 2025/056713 A1: EGFR Antagonists for the treatment of diseases involving unwanted migration, proliferation, and metaplasia of retinal pigment epithelium (RPE) cells. A.P.K has acted as a paid consultant or lecturer to Abbvie, Aerie, Google Health, Heidelberg Engineering, Glaucore, Novartis, Qlaris Bio, Regeneron, Reichert, Santen, Seonix Bio, Tavo Biotherapeutics, Thea and Topcon. The remaining authors declare no competing interests.

Funding

This work was supported by the Engineering and Physical Sciences Research Council [grant number UKRI1662].

Author Contribution

H.K., M.Gr., and R.A.J. designed the study, implemented the training and evaluation pipeline, and performed the statistical analyses. M.Gr., R.A.J., and S.H. performed annotation. A.A., D.C.A., A.K.D., S.Ki., and B.P. contributed to study conceptualization. L.J., Y.L., J.Hu., K.I., P.N., K.H., M.G.M., T.E., YI.W., T.N., X.D., A.Y.O., S.C.E., C.L., and A.L. contributed to data preprocessing, model development, validation, and manuscript writing. J.C., D.L., H.Ki., and Y.S. collected and

managed datasets. A.P.K., R.L., A.W., Y.Wa., S.P.J., J.B.J., M.M.B., G.M.K., and V.N. managed cohort data and provided clinical expertise. S.K.W., Y.Z., S.H., and P.A.K. supervised the project. All authors contributed to the writing of the manuscript and approved its final version.

Acknowledgement

This work was supported by the Engineering and Physical Sciences Research Council [grant number UKRI1662] and by the National Institute for Health and Care Research (NIHR) Biomedical Research Centre based at Moorfields Eye Hospital National Health Service (NHS) Foundation Trust and University College London (UCL) Institute of Ophthalmology. The views expressed are those of the author(s) and not necessarily those of the NHS, NIHR, or Department of Health and Social Care. M.Gr. is supported by the Dr. Werner Jackstädt-Foundation.

Data Availability

The publicly available datasets used in this study can be accessed as follows: ADAM (<https://amd.grand-challenge.org/>; also available via IEEE DataPort at <https://doi.org/10.21227/dt4f-rt59>) and IDRid (<https://idrid.grand-challenge.org/>). The UEMS, CIEMS, and BES datasets are not publicly available due to participant privacy and institutional data-sharing agreements, but de-identified data may be made available from the corresponding investigators (M.M.B./G.M.K., V.N., and J.B.J./Y.Wa., respectively) upon reasonable request and approval by the relevant institutional review boards. The evaluation framework and datasets curated in this study, comprising the Hanyang University Guri Hospital datasets (HYU-DR-CFP, HYU-AMD-CFP, and HYU-AMD-OCT) and the Yonsei-AMD-OCT dataset, are available at https://github.com/hyunminkim0625/PointGuided_Benchmark under a CC BY-NC 4.0 license. The publicly available foundation models evaluated in this study can be obtained from their respective repositories: RETFound-DINOv2 and RETFound-MAE (<https://github.com/rmaphoh/RETFound>), VisionFM-CFP and VisionFM-OCT (<https://github.com/ABILab-CUHK/VisionFM>), RETFound-Green (https://github.com/justinengelm/RETFound_Green), RetiZero (<https://github.com/LooKing9218/RetiZero>), UrFound (<https://github.com/yukkai/UrFound>), MIRAGE (<https://github.com/j-morano/MIRAGE>), DINOv3 (<https://github.com/facebookresearch/dinov3>), Perception Encoder (https://github.com/facebookresearch/perception_models), and Efficient Universal Perception Encoder (<https://github.com/facebookresearch/EUPE>). Image quality control was performed using the automated retinal image analysis tool AutoMorph (<https://github.com/rmaphoh/AutoMorph>).

References

1. Siméoni, O. et al. *Dinov3*. *arXiv preprint arXiv:2508.10104* (2025).
2. Oquab, M. et al. *Dinov2: Learning Robust Visual Features without Supervision*. *arXiv preprint arXiv:2304.07193* (2023).
3. Caron, M. et al. in *Proceedings of the IEEE/CVF International Conference on Computer Vision* 9650–9660 (2021).
4. Bolya, D. et al. Perception encoder: the best visual embeddings are not at the output of the network. *Advances in Neural Information Processing Systems* 38, 60884–60937 (2025).
5. Zhu, C. et al. *Efficient Universal Perception Encoder*. *arXiv preprint arXiv:2603.22387* (2026).
6. 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., UK Biobank Eye & Vision Consortium, Altmann, A., Lee, A. Y., Topol, E. J., Denniston, A. K., Alexander, D. C., & Keane, P. A. A foundation model for generalizable disease detection from retinal images. *Nature* 622, 156–163 (2023).
7. Zhou, Y. et al. *Generalist versus Specialist Vision Foundation Models for Ocular Disease and Oculomics*. *arXiv preprint arXiv:2509.03421* (2025).
8. Yu, K. et al. in *International Conference on Medical Image Computing and Computer-Assisted Intervention* 753–762 (Springer. 2024).
9. Liu, Z., Xu, H., Woicik, A., Shapiro, L. G., Blazes, M., Wu, Y., Steffen, V., Cukras, C. A., Lee, C. S., Zhang, M., Lee, A. Y., & Wang, S. A three-dimensional multi-modal foundation model for optical coherence tomography. *Nature Biomedical Engineering* (2026).
10. Wang, M., Lin, T., Lin, A., Yu, K., Peng, Y., Wang, L., Chen, C., Zou, K., Liang, H., Chen, M., Yao, X., Zhang, M., Huang, B., Zheng, C., Zhang, P., Chen, W., Luo, Y., Chen, Y., Xia, H., Shi, T., Zhang, Q., Guo, J., Chen, X., Wang, J., Tham, Y. C., Liu, D., Wong, W., Thakur, S., Fenner, B. J., Fang, D., Liu, S., Liu, Q., Huang, Y., Zeng, H., Meng, Y., Zhou, Y., Jiang, Z., Qiu, M., Zhang, C., Chen, X., Wang, S. Y., Lee, C. S., Sobrin, L., Cheung, C. Y., Pang, C. P., Keane, P. A., Cheng, C. Y., Chen, H., & Fu, H. Enhancing diagnostic accuracy in rare and common fundus diseases with a knowledge-rich vision-language model. *Nature Communications* 16, 5528 (2025).
11. Morano, J., Fazekas, B., Sükei, E., Fecso, R., Emre, T., Gumpinger, M., Faustmann, G., Oghbaie, M., Schmidt-Erfurth, U., & Bogunović, H. Multimodal foundation model and benchmark for comprehensive retinal OCT image analysis. *NPJ Digital Medicine* 8, 576 (2025).
12. Wilkinson, C. P., Ferris, F. L., Klein, R. E., Lee, P. P., Agardh, C. D., Davis, M., Dills, D., Kampik, A., Pararajasegaram, R., Verdaguer, J. T., & Global Diabetic Retinopathy Project Group Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology* 110, 1677–1682 (2003).
13. Dai, L., Wu, L., Li, H., Cai, C., Wu, Q., Kong, H., Liu, R., Wang, X., Hou, X., Liu, Y., Long, X., Wen, Y., Lu, L., Shen, Y., Chen, Y., Shen, D., Yang, X., Zou, H., Sheng, B., & Jia, W. A deep learning system for detecting diabetic retinopathy across the disease spectrum. *Nature Communications* 12, 3242 (2021).
14. Schlegl, T., Waldstein, S. M., Bogunovic, H., Endstraßer, F., Sadeghipour, A., Philip, A. M., Podkowinski, D., Gerendas, B. S., Langs, G., & Schmidt-Erfurth, U. Fully automated detection and quantification of macular fluid in OCT using deep learning. *Ophthalmology* 125, 549–558 (2018).
15. Schmidt-Erfurth, U., Chong, V., Loewenstein, A., Larsen, M., Souied, E., Schlingemann, R., Eldem, B., Monés, J., Richard, G., Bandello, F., & European Society of Retina Specialists. Guidelines for the management of neovascular age-related macular degeneration by the European Society of Retina Specialists

- (EURETINA). *The British Journal of Ophthalmology* 98, 1144–1167 (2014).
16. De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., Van Den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., Hughes, C. O., Raine, R., Hughes, J., Sim, D. A., Egan, C., Tufail, A., Montgomery, H., Hassabis, D., Rees, G., Back, T., Khaw, P. T., Suleyman, M., Cornebise, J., Keane, P. A., & Ronneberger, O. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine* 24, 1342–1350 (2018).
 17. Rim, T. H., Lee, A. Y., Ting, D. S., Teo, K. Y. C., Yang, H. S., Kim, H., Lee, G., Teo, Z. L., Teo Wei Jun, A., Takahashi, K., Yoo, T. K., Kim, S. E., Yanagi, Y., Cheng, C. Y., Kim, S. S., Wong, T. Y., & Cheung, C. M. G. Computer-aided detection and abnormality score for the outer retinal layer in optical coherence tomography. *The British Journal of Ophthalmology* 106, 1301–1307 (2022).
 18. Wu, Y., Qian, B., Li, T., Qin, Y., Guan, Z., Chen, T., Jia, Y., Zhang, P., Zeng, D., Moroi, S., Raman, R., Thinggaard, B. S., Pedersen, F., Nehe, J. A. O., Kamalden, T. A., Zhou, Y., Jin, Y., Li, H., Ran, A. R., Yang, D., Meng, Z., Peng, Q., Zheng, Y. F., Wang, D., Ji, H., Zang, P., Yin, C., Shen, J., Chen, Y., Yu, W., Dai, R., Zhang, C., Zhao, X., Wang, X., Chen, Y., Wu, Q., Xie, H., Szeto, S. K. H., Chan, J. Y. Y., Chan, V. T. T., Xie, H. T., Wei, R., Li, J., Ma, W., Zhu, L., Wang, H., Fu, H., Wang, W., Lin, S., Xu, Z., Guan, N., Zhang, X., Grzybowski, A., Gołębiewska-Bogaj, M., Gawęcki, M., Smedowski, A., Szaraniec, W., Wu, Y., Wen, Y., Chen, X., Yao, Y., EyeFM Global Reader Study Team, Lim, L. L., Cheung, C. Y., Tan, G. S. W., Grauslund, J., Ruamviboonsuk, P., Sivaprasad, S., Keane, P. A., Wang, Y. X., Tham, Y. C., Cheng, C. Y., Wong, T. Y., & Sheng, B. An eyecare foundation model for clinical assistance: a randomized controlled trial. *Nature Medicine* 31, 3404–3413 (2025).
 19. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., & Collins, R. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine* 12, e1001779 (2015).
 20. Bikbov, M., Fayzrakhmanov, R. R., Kazakbaeva, G., & Jonas, J. B. Ural Eye and Medical Study: description of study design and methodology. *Ophthalmic Epidemiology* 25, 187–198 (2018).
 21. Jonas, J. B., Xu, L., & Wang, Y. X. The Beijing Eye study. *Acta Ophthalmologica* 87, 247–261 (2009).
 22. Engelmann, J., & Bernabeu, M. O. Training a high-performance retinal foundation model with half-the-data and 400 times less compute. *Nature Communications* 16, 6862 (2025).
 23. Qiu, J. et al. *Visionfm: a Multi-modal Multi-task Vision Foundation Model for Generalist Ophthalmic Artificial Intelligence*. *arXiv preprint arXiv:2310.04992* (2023).
 24. Baharoon, M. et al. *Evaluating General Purpose Vision Foundation Models for Medical Image Analysis: an Experimental Study of dinov2 on Radiology Benchmarks*. *arXiv preprint arXiv:2312.02366* (2023).
 25. Isztli, D., Spitznagel, T., Somfai, G. M., & Santos, R. *When Do Domain-Specific Foundation Models Justify Their Cost? A Systematic Evaluation Across Retinal Imaging Tasks*. *arXiv preprint arXiv:2511.22001* (2025).
 26. Nouyed, M. I., Al-Mamun, M., Adjero, D. A., & Hu, G. Comparative Analysis of General-Purpose vs. Domain-Specific Multimodal Models for Diabetic Retinopathy Classification. *Diagnostics* 16, 1504 (2026).
 27. Hemelings, R., Elen, B., Schuster, A. K., Blaschko, M. B., Barbosa-Breda, J., Hujanen, P., Junglas, A., Nickels, S., White, A., Pfeiffer, N., Mitchell, P., De Boever, P., Tuulonen, A., & Stalmans, I. A generalizable deep learning regression model for automated glaucoma screening from fundus images. *npj Digital Medicine* 6, 112 (2023).
 28. DeGrave, A. J., Janizek, J. D., & Lee, S. I. AI for radiographic COVID-19 detection selects shortcuts over signal. *medRxiv*. (2020).
 29. Kumar, A., Raghunathan, A., Jones, R., Ma, T., & Liang, P. *Fine-Tuning Can Distort Pretrained Features and Underperform Out-of-Distribution*. *arXiv preprint arXiv:2202.10054* (2022).
 30. Hegde, N., Hipp, J. D., Liu, Y., Emmert-Buck, M., Reif, E., Smilov, D., Terry, M., Cai, C. J., Amin, M. B., Mermel, C. H., Nelson, P. Q., Peng, L. H., Corrado, G. S., & Stumpe, M. C. Similar image search for histopathology: SMILY. *NPJ Digital Medicine* 2, 56 (2019).
 31. Liu, Z., Xu, H., Woicik, A., Shapiro, L. G., Blazes, M., Wu, Y., Steffen, V., Cukras, C. A., Lee, C. S., Zhang, M., Lee, A. Y., & Wang, S. A three-dimensional multi-modal foundation model for optical coherence tomography. *Nature Biomedical Engineering* (2026).
 32. Maron, R. C., Hekler, A., Kriehoff-Henning, E., Schmitt, M., Schlager, J. G., Utikal, J. S., & Brinker, T. J. Reducing the impact of confounding factors on skin cancer classification via image segmentation: technical model study. *Journal of Medical Internet Research* 23, e21695 (2021).
 33. Winkler, J. K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., & Haenssle, H. A. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatology* 155, 1135–1141 (2019).
 34. Koh, P. W. et al. in International conference on machine learning 5338–5348 (PMLR, 2020).
 35. Bearman, A., Russakovsky, O., Ferrari, V., & Fei-Fei, L. in European conference on computer vision 549–565 (Springer, 2016).
 36. Dosovitskiy, A. et al. *An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale*. *arXiv preprint arXiv:2010.11929* (2020).
 37. Ferris, F. L., 3rd, Wilkinson, C. P., Bird, A., Chakravarthy, U., Chew, E., Csaky, K., Sadda, S. R., & Beckman Initiative for Macular Research Classification Committee Clinical classification of age-related macular degeneration. *Ophthalmology* 120, 844–851 (2013).
 38. Fang, H., Li, F., Fu, H., Sun, X., Cao, X., Lin, F., Son, J., Kim, S., Quellec, G., Matta, S., Shankaranarayana, S. M., Chen, Y. T., Wang, C. H., Shah, N. A., Lee, C. Y., Hsu, C. C., Xie, H., Lei, B., Baid, U., Innani, S., Dang, K., Shi, W., Kamble, R., Singhal, N., Wang, C. W., Lo, S. C., Orlando, J. I., Bogunovic, H., Zhang, X., & Xu, Y. ADAM challenge: detecting age-related macular degeneration from fundus images. *IEEE Transactions on Medical Imaging* 41, 2828–2847 (2022).
 39. Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabudhe, V., & Meriaudeau, F. Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. *Data* 3, 25 (2018).

40. Nangia, V., Matin, A., Bhojwani, K., Kulkarni, M., Yadav, M., & Jonas, J. B. Optic disc size in a population-based study in central India: the Central India Eye and Medical Study (CIEMS). *Acta Ophthalmologica* 86, 103–104 (2008).
41. Zhou, Y., Wagner, S. K., Chia, M. A., Zhao, A., Woodward-Court, P., Xu, M., Struyven, R., Alexander, D. C., & Keane, P. A. AutoMorph: automated retinal vascular morphology quantification via a deep learning pipeline. *Translational Vision Science & Technology* 11, 12 (2022).
42. Warwick, A. N., Curran, K., Hamill, B., Stuart, K., Khawaja, A. P., Foster, P. J., Lotery, A. J., Quinn, M., Madhusudhan, S., Balaskas, K., Peto, T., & UKBB Eye and Vision Consortium UK Biobank retinal imaging grading: methodology, baseline characteristics and findings for common ocular diseases. *Eye* 37, 2109–2116 (2023).
43. Loshchilov, I., & Hutter, F. *Decoupled Weight Decay Regularization*. *arXiv preprint arXiv:1711.05101* (2017).
44. Wang, H., Jie, S., & Deng, Z. Focus your attention when few-shot classification. *Advances in Neural Information Processing Systems* 36, 59689–59707 (2023).
45. Zhu, M. et al. *Attention-Guided Fair AI Modeling for Skin Cancer Diagnosis*. *arXiv preprint arXiv:2511.21775* (2025).
46. Khattab, O., & Zaharia, M. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 39–48 (2020).

Figures

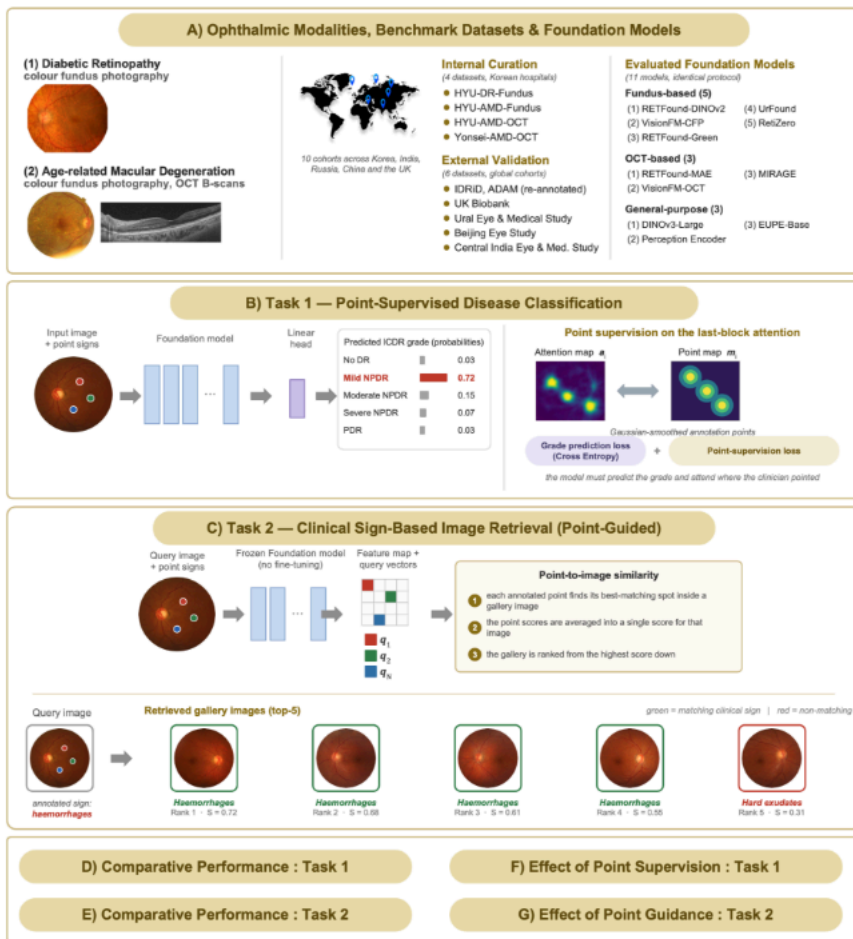


Figure 1

Framework overview

(A) Study overview. Two ophthalmic diseases, diabetic retinopathy (DR) and age-related macular degeneration (AMD), were studied across two imaging modalities, colour fundus photography (CFP) and optical coherence tomography (OCT). Eleven foundation models (five fundus-based, three OCT-based and three general-purpose) were evaluated under an identical protocol on four internally curated datasets and six established datasets, spanning 10 cohorts from Korea, India, Russia, China and the UK. (B) Task 1, point-supervised disease classification, in which a point-supervision loss aligns the last-block attention map with the Gaussian-smoothed clinician annotation points and is optimised jointly with the grade-prediction (cross-entropy) loss. (C) Task 2, point-guided clinical sign-based image retrieval. A frozen foundation model encodes each image without fine-tuning. Patches whose point-derived weight exceeds uniform are taken as query patches, and gallery images are ranked by the weighted average of query patch similarities, where the weights reflect the spatial density of the annotated points. (D, E) Comparative performance of all 11 models on Task 1 and Task 2, respectively. (F, G) Effect of point supervision on Task

1 (F) and of point guidance on Task 2 (G), expressed as the change relative to the corresponding naive baseline. The fundus and OCT images, point annotations and retrieval results shown are illustrative examples only and do not represent actual data or model output.

Abbreviations: ADAM, Automatic Detection challenge on Age-related Macular degeneration; AMD, age-related macular degeneration; BES, Beijing Eye Study; CFP, colour fundus photograph; CIEMS, Central India Eye and Medical Study; DR, diabetic retinopathy; HYU, Hanyang University; ICDR, International Clinical Diabetic Retinopathy; IDRiD, Indian Diabetic Retinopathy Image Dataset; NPDR, non-proliferative diabetic retinopathy; OCT, optical coherence tomography; PDR, proliferative diabetic retinopathy; UEMS, Ural Eye and Medical Study.

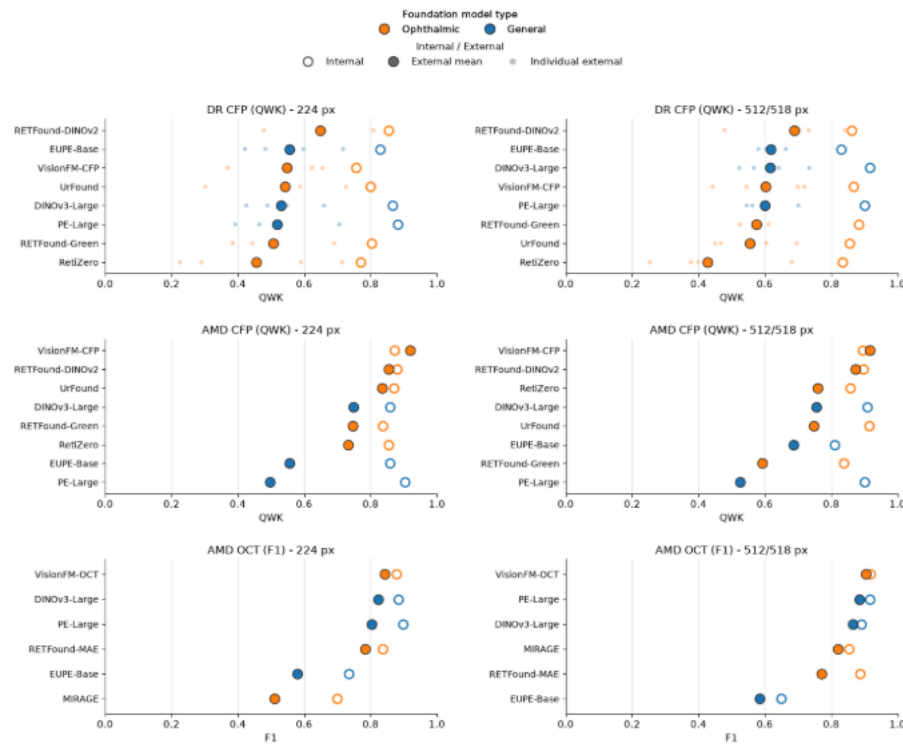


Figure 2

Task 1 – Point-supervised classification: DR and AMD in fundus and OCT modalities.

Foundation model performance across three settings: DR grading on CFP (5-class quadratic weighted kappa, QWK), AMD grading on CFP (4-class QWK), and AMD classification on OCT (3-class macro-F1). Results are shown separately for the 224 px (left column) and 512/518 px (right column) settings. Unfilled circles denote internal validation and filled circles denote external validation. For DR-CFP, where four external datasets were available, the filled circle is the macro-average across the four datasets and the small translucent points are the individual dataset scores; the remaining settings have a single external dataset. Orange denotes ophthalmic foundation models and blue denotes general-purpose foundation models. Within each panel, models are ordered by external performance. Per-dataset scores with 95% bootstrap confidence intervals (N = 2,000 resamples, performed at the patient level except for IDRiD and ADAM, which lack patient identifiers) are provided in Supplementary Figures 2-4 for both internal and external validation.

Abbreviations: ADAM, Automatic Detection challenge on Age-related Macular degeneration; AMD, age-related macular degeneration; BES, Beijing Eye Study; CFP, colour fundus photograph; CIEMS, Central India Eye and Medical Study; DR, diabetic retinopathy; ICDR, International Clinical Diabetic Retinopathy; IDRiD, Indian Diabetic Retinopathy Image Dataset; OCT, optical coherence tomography; QWK, quadratic weighted kappa; UEMS, Ural Eye and Medical Study.

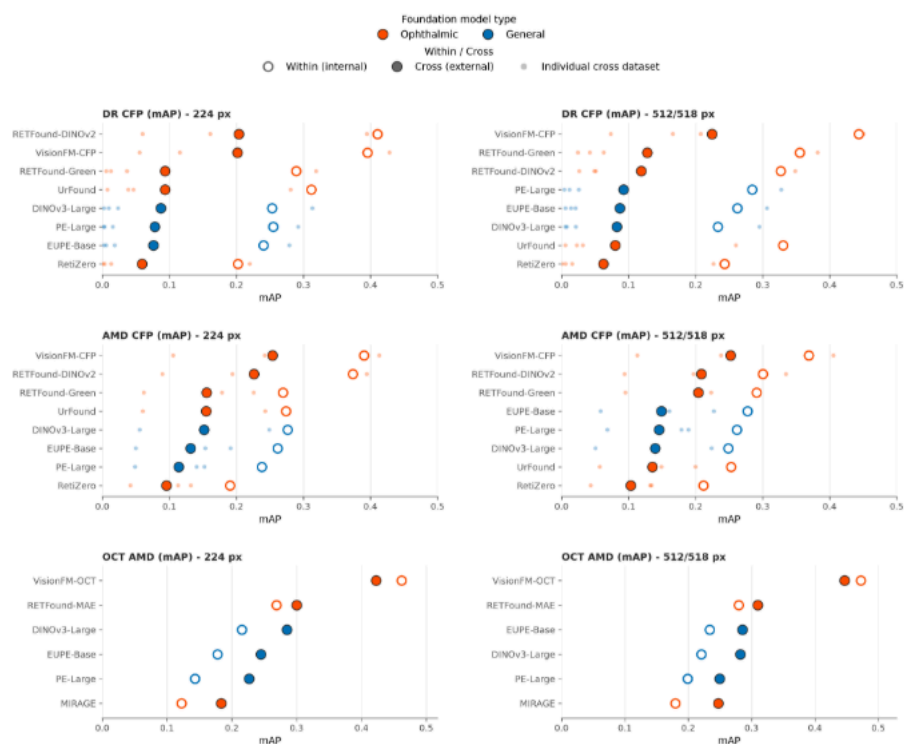


Figure 3

Task 2 — Point-guided retrieval: DR and AMD in fundus and OCT modalities.

Retrieval performance under point guidance, reported as mean average precision (mAP), for diabetic retinopathy (DR) and age-related macular degeneration (AMD) in colour fundus photography (CFP) and for AMD in optical coherence tomography (OCT). Results are shown separately for the 224 px (left column) and 512/518 px (right column) settings. mAP is the unweighted average across clinical sign categories within each dataset. Where more than one dataset was available, these per-dataset values were then averaged once more across datasets, again unweighted. Unfilled circles denote within-dataset retrieval (internal validation) and filled circles denote cross-dataset retrieval (external validation); for the CFP settings, where multiple external datasets were available, the small translucent points are the per-dataset values and the filled circle is their unweighted mean, whereas the OCT setting has a single external dataset. Orange denotes ophthalmic foundation models and blue denotes general-purpose foundation models. Within each panel, models are ordered by cross-dataset performance. Per-sign and per-dataset scores with 95% bootstrap confidence intervals (N = 2,000 resamples of the query set, drawn at the patient level) are provided in Supplementary Figures 5-7.

Abbreviations: AMD, age-related macular degeneration; CFP, colour fundus photograph; DR, diabetic retinopathy; mAP, mean average precision; OCT, optical coherence tomography.

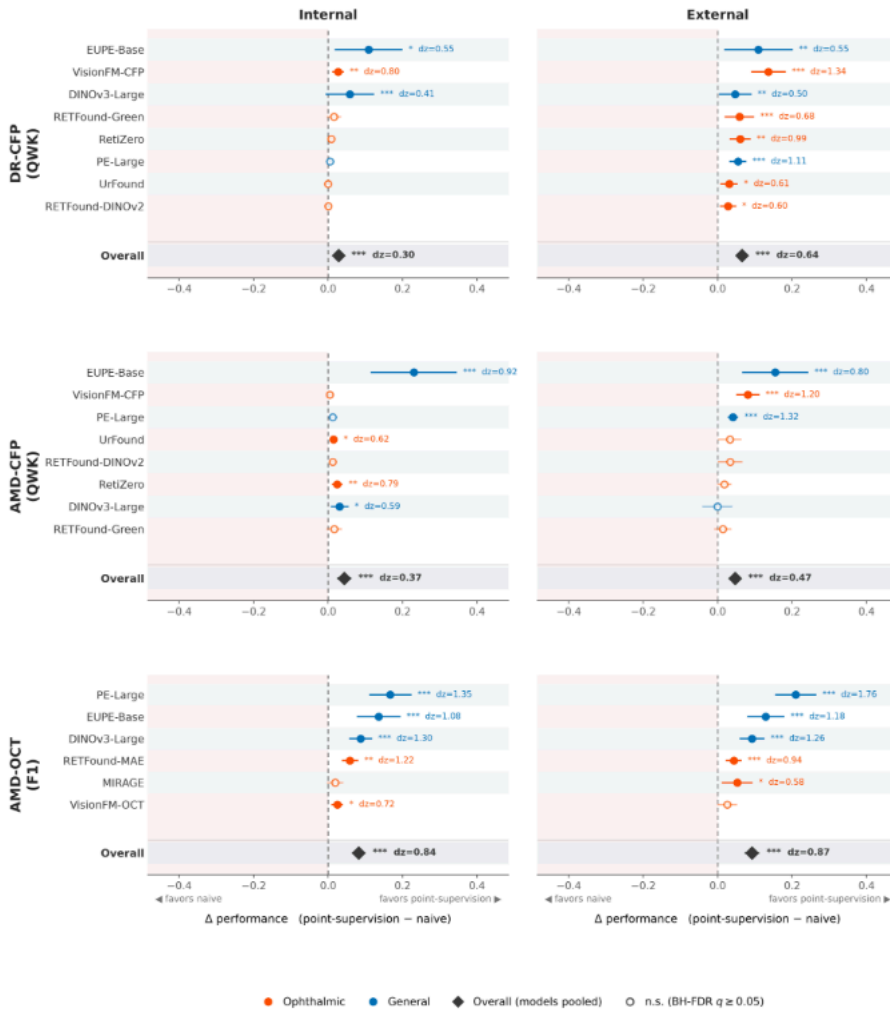


Figure 4

Task 1: Effect of point supervision on disease classification performance.

Forest plot of the paired change in classification performance (Δ = point-supervision - naive) for DR grading on colour fundus photographs (DR-CFP; QWK), AMD grading on CFP (AMD-CFP; QWK) and AMD classification on OCT (AMD-OCT; macro-F1). Rows of panels correspond to these three settings and columns to internal and external validation. Circles show the mean Δ for each foundation model, averaged across all hyperparameter configurations and input resolutions, not at the best-performing configuration; black diamonds ("Overall") pool all models within a setting. Error bars are 95% confidence intervals of the mean Δ . Positive values, to the right of the dashed line at zero, favour point supervision over naive training. Orange denotes ophthalmic foundation models and blue denotes general-purpose foundation models. Significance was assessed with the Wilcoxon signed-rank test on the paired runs and corrected across all comparisons using the Benjamini–Hochberg false discovery rate; the resulting adjusted p values are denoted q. Filled markers are significant and unfilled markers are non-significant ($q \geq 0.05$). * $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$. dz is the paired effect size (Cohen's dz), the mean of the paired differences divided by their standard deviation, and is shown only for significant comparisons.

Abbreviations: AMD, age-related macular degeneration; BH, Benjamini–Hochberg; CFP, colour fundus photograph; dz, paired Cohen's d; FDR, false discovery rate; DR, diabetic retinopathy; OCT, optical coherence tomography; QWK, quadratic weighted kappa.

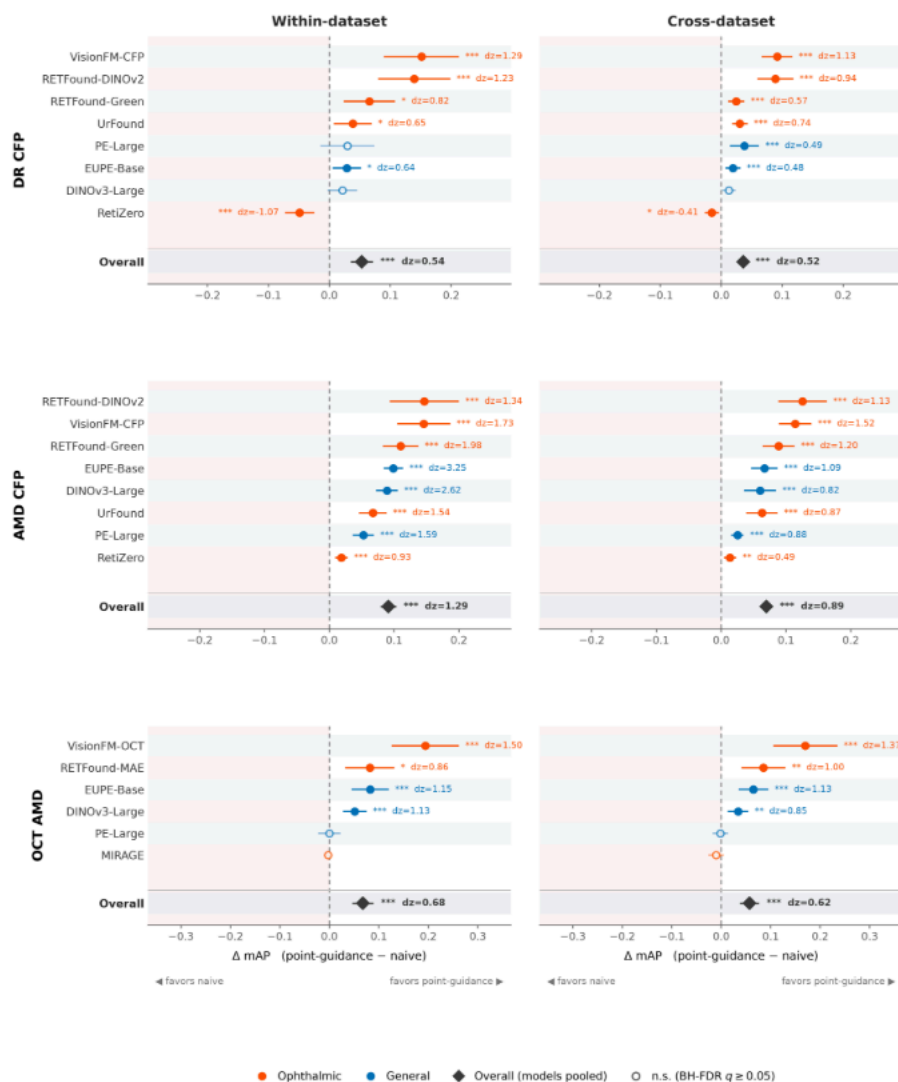


Figure 5

Task 2: Effect of point guidance on retrieval performance for each foundation model.

Forest plot of the paired change in retrieval performance (Δ mAP = point-guidance - naive) for each foundation model, shown for DR-CFP, AMD-CFP and AMD-OCT. Rows of panels correspond to these three settings and columns to within-dataset and cross-dataset evaluation; within each panel the y-axis lists the individual models, with a bottom "Overall" row pooling all models. For each model, Δ was computed per matched (clinical sign $\times$ dataset $\times$ input resolution) pair. Circles show the mean Δ per model and black diamonds show the pooled estimate across models; error bars are 95% confidence intervals. Positive values, to the right of the dashed line at zero, favour point guidance over naive retrieval. Orange denotes ophthalmic foundation models and blue denotes general-purpose foundation models. Significance was assessed with the Wilcoxon signed-rank test and corrected across all comparisons using the Benjamini-Hochberg false discovery rate; the resulting adjusted p values are denoted q. Filled markers are significant and unfilled markers are non-significant ($q \geq 0.05$). * $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$. For significant comparisons, Cohen's dz (mean Δ divided by the SD of Δ across matched pairs) is annotated next to each marker. Within each panel, models are ordered by within-dataset Δ .

Abbreviations: AMD, age-related macular degeneration; CFP, colour fundus photograph; DR, diabetic retinopathy; dz, paired Cohen's d; mAP, mean average precision; OCT, optical coherence tomography; SD, standard deviation.



Figure 6

Task 2: Effect of point guidance on retrieval performance for each clinical sign.

Forest plot of the paired change in retrieval performance (Δ mAP = point-guidance - naive), shown per clinical sign for DR-CFP, AMD-CFP and AMD-OCT. Rows of panels correspond to these three settings and columns to within-dataset (internal) and cross-dataset (external) evaluation; within each panel the y-axis lists the individual clinical signs, with a bottom "Overall" row pooling all signs. For each sign, Δ was computed per matched (model $\times$ dataset $\times$ input resolution) pair. Circles show the mean Δ per sign and diamonds show the pooled estimate across signs; error bars are 95% confidence intervals. Positive values, to the right of the dashed line at zero, favour point guidance over naive retrieval. Significance was assessed with the Wilcoxon signed-rank test and corrected across all comparisons using the Benjamini-Hochberg false discovery rate; the resulting adjusted p values are denoted q. Filled markers are significant and unfilled markers are non-significant ($q \geq 0.05$). * $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$. For significant comparisons, Cohen's d (mean Δ divided by the SD of Δ across matched pairs) is annotated next to each marker.

Abbreviations: AMD, age-related macular degeneration; BH, Benjamini-Hochberg; CFP, colour fundus photograph; CNV, choroidal neovascularization; DR, diabetic retinopathy; dz, paired Cohen's d; FDR, false discovery rate; IRF, intraretinal fluid; IRMA, intraretinal microvascular abnormality; mAP, mean average precision; OCT, optical coherence tomography; PED, pigment epithelial detachment; RPE, retinal pigment epithelium; SD, standard deviation; SHRM, subretinal hyperreflective material; SRF, subretinal fluid.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [pointguidedhyunminkimsupple.docx](#)