# Supplementary Information: UWF-FM: A Foundation Model for Comprehensive Ultra-Widefield Retinal Assessment

## 1 Category-level performance comparison across clinically structured retinal assessment settings

To determine whether the observed improvements were broadly distributed across retinal assessment settings rather than driven by a limited subset of categories, we compared category-level AUC values between UWF-FM and alternative pretrained backbones under both frozen-feature and full fine-tuning settings.

Across all clinically structured retinal assessment settings, including disease diagnosis, DR assessment, and retinal finding characterization, UWF-FM demonstrated consistently higher AUC values across the majority of categories under both full fine-tuning and frozen-feature evaluation (Fig. S1). Performance differences were particularly pronounced under frozen-feature evaluation and during cross-vendor transfer to Zeiss images, where non-UWF pretrained models exhibited larger reductions in category-level performance.

Overall, these category-level analyses indicate that the benefits of UWF pretraining are broadly distributed across retinal assessment settings and are not driven by a limited subset of diseases or retinal findings.

## 2 Additional external validation of diabetic retinopathy assessment in an independent Shanghai cohort

To further evaluate the robustness of UWF-FM for diabetic retinopathy assessment, we performed an additional external validation using an independent cohort of 198 UWF images collected from clinical centers in Shanghai, China. Consistent with the findings observed in the primary external DR cohort, UWF-FM achieved an mAUC of 0.907, outperforming foundation models pretrained on non-UWF datasets (Table S1). The reproducibility of these performance gains across two independent external DR cohorts further supports the robustness and generalizability of UWF-specific pretraining across geographically distinct clinical environments.

**Table S1 Additional external validation of diabetic retinopathy assessment in External-Optos-DR1.** Results are reported as mAUC (mean ± s.d.) using the MLP fine-tuning protocol.

| | DINOV3-LVD | MAE-RETFound | MAE-In1K | UWF-FM |
|---|---|---|---|---|
| mAUC | $0.570 \pm 0.083$ | $0.654 \pm 0.043$ | $0.840 \pm 0.043$ | $\mathbf{0.907 \pm 0.034}$ |

# 3 Additional evaluation under clinically relevant operating thresholds

To further characterize model behavior under clinically relevant decision conditions, we evaluated performance across a broader range of fixed sensitivity and fixed specificity thresholds.

Across all clinically structured retinal assessment settings, UWF-FM consistently achieved higher sensitivity than comparison models at fixed specificity thresholds ranging from 80% to 95% (Fig. S2). Similarly, at fixed sensitivity thresholds of 90% and 95%, UWF-FM achieved higher specificity while maintaining high sensitivity. These advantages were preserved across disease diagnosis, DR assessment, retinal finding characterization, and cross-vendor evaluation.

Category-level analyses further demonstrated that the advantages of UWF-FM were broadly distributed across clinically structured retinal assessment settings rather than being restricted to a limited subset of categories (Fig S3, Fig. S4, Fig. S5, and Fig S6). Although the magnitude of improvement varied across categories, most disease, DR assessment, and retinal finding categories exhibited more favorable operating characteristics for UWF-FM than for comparison models.

Notably, strong performance was also observed for several retinal findings requiring urgent referral, including retinal detachment, retinal breaks, vitreous hemorrhage, and retinal neovascularization. Many of these lesions arise predominantly in the peripheral retina or present with heterogeneous morphology, suggesting that UWF-specific pretraining may provide particular advantages for clinically important peripheral retinal pathology.

# 4 Additional analyses of AI-assisted clinician interpretation

To further characterize the heterogeneity of AI-assisted performance observed in the main analysis, we performed additional analyses across retinal finding categories and individual clinicians.

## 4.1 Retinal finding-level heterogeneity of AI-assisted interpretation

We examined the effects of AI assistance across individual retinal findings by analyzing retinal finding-level receiver operating characteristic (ROC) curves and corresponding

clinician operating points (Fig. S7). Substantial variability was observed across retinal findings in both the magnitude and direction of performance changes. For some conditions, AI assistance shifted clinician operating points towards higher sensitivity with improved overall performance, whereas for others, changes were minimal or reflected trade-offs between sensitivity and specificity. These findings are consistent with differences in baseline task difficulty, lesion characteristics, and data distribution across retinal conditions.

### 4.2 Reader-level and within-reader heterogeneity

To further assess how AI assistance influences individual clinicians, we analyzed performance changes separately for each clinician across retinal finding categories (Fig. S8).

At the reader level, the magnitude of performance changes differed across clinicians. While all readers exhibited similar overall trends, improvements were more pronounced in some junior clinicians, whereas others demonstrated more modest gains.

Importantly, substantial variability was also observed within individual readers across different retinal findings. For a given clinician, AI assistance resulted in performance improvements for certain retinal findings while having limited or mixed effects for others. This indicates that the impact of AI assistance is not uniform even within the same reader, but depends on the specific diagnostic context.

Taken together, these analyses demonstrate multi-level heterogeneity in AI-assisted performance, spanning retinal finding-specific differences, variability across clinicians, and context-dependent responses within individual readers.

### 4.3 Mixed-effects analysis of clinician-assisted interpretation

To further account for repeated reader-case observations in the crossover clinician study, mixed-effects logistic regression models were fitted separately for sensitivity and specificity to account for repeated reader-case observations (Shown in Table S2). For sensitivity analysis, only positive reference labels were included, whereas specificity analysis included only negative reference labels.

AI assistance condition (unaided versus AI-assisted) was included as a fixed effect. Reader identity and image case were modeled as random intercepts to account for inter-reader variability and repeated case interpretation across clinicians. Effect sizes were summarized using odds ratios (ORs), with corresponding two-sided $p$-values reported. Each observation corresponded to a binary reader interpretation for a single retinal finding on a given image. An OR greater than 1 indicates that AI assistance increased the odds of a correct clinician response for the corresponding endpoint, whereas an OR less than 1 indicates reduced odds.

## 5 Details of the dataset

A detailed breakdown of the downstream datasets used in this study is provided below. Datasets are organized according to the clinically structured retinal assessment settings used throughout the manuscript, including disease diagnosis,

**Table S2 Mixed-effects analysis of AI-assisted clinician interpretation.**

| Endpoint | Odds ratio (OR) | Standard error | Wald $z$-statistic | $P$ value | Observations |
|---|---|---|---|---|---|
| Sensitivity | 1.45 | 0.035 | 10.62 | $< 0.001$ | 11,416 |
| Specificity | 0.87 | 0.020 | $-6.98$ | $< 0.001$ | 122,984 |

**Table S3** Distribution of retinal finding categories across the internal development cohort (Optos-UWF-Lesion), cross-vendor cohort (Zeiss-UWF-Lesion), external validation cohort (External-Optos-Lesion), and clinician evaluation cohort (Clinical-Optos-Lesion).

| **Lesion category** | **Optos-UWF** | **Zeiss-UWF** | **External-Optos** | **Clinical-Optos** |
|---|---|---|---|---|
| Refractive media opacity | 2467 | 155 | 107 | 24 |
| Vitreous hemorrhage | 426 | 47 | 32 | 38 |
| Asteroid hyalosis | 87 | 26 | 14 | 50 |
| Posterior vitreous detachment | 256 | 32 | 37 | 33 |
| Other vitreous abnormalities | 677 | 40 | 61 | 25 |
| Drusen | 347 | 1313 | 18 | 44 |
| Myopic macular atrophy | 212 | 270 | 29 | 43 |
| Epiretinal membrane | 97 | 389 | 4 | 41 |
| Blurred disc margin | 105 | 35 | 2 | 41 |
| Large cup-to-disc ratio | 330 | 25 | 4 | 35 |
| Optic nerve atrophy | 156 | 461 | 18 | 26 |
| High myopic disc atrophy | 3278 | 837 | 140 | 67 |
| Retinal detachment | 607 | 84 | 45 | 37 |
| Subretinal hemorrhage | 80 | 12 | 15 | 31 |
| Preretinal hemorrhage | 442 | 40 | 28 | 36 |
| Retinal break | 450 | 55 | 36 | 33 |
| Retinal fibrous membrane | 460 | 65 | 10 | 42 |
| Retinal laser scars | 1518 | 281 | 108 | 30 |
| Panretinal photocoagulation | 468 | 391 | 109 | 41 |
| Bone spicule pigmentation | 94 | 109 | 9 | 50 |
| Leopard fundus | 4171 | 1901 | 191 | 28 |
| Retinal hemorrhages / spots | 4087 | 1696 | 191 | 11 |
| Hard exudates | 2677 | 1112 | 54 | 14 |
| Cotton-wool spots | 1792 | 722 | 32 | 9 |
| Retinal neovascularization | 526 | 65 | 18 | 29 |
| Old retinal pigment lesions | 430 | 356 | 18 | 29 |
| Peripheral retinal degeneration | 1364 | 73 | 47 | 37 |
| Retinal vascular sheathing | 639 | 1073 | 50 | 7 |

diabetic retinopathy (DR) assessment, retinal finding characterization, and tumor classification. Detailed category distributions are summarized in Tables S3–S7.

**Table S4** Distribution of disease diagnosis categories across the internal development cohort (Optos-UWF-Dis), cross-vendor cohort (Zeiss-UWF-Dis), and external validation cohort (External-Optos-Dis).

| Disease category | Optos-UWF | Zeiss-UWF | External-Optos |
|---|---|---|---|
| Normal fundus | 416 | 41 | 223 |
| Retinitis pigmentosa | 156 | 141 | 63 |
| Central retinal vein occlusion | 123 | 33 | 25 |
| Branch retinal vein occlusion | 176 | 78 | 6 |
| Familial exudative vitreoretinopathy | 45 | 10 | 8 |
| Non-proliferative diabetic retinopathy | 834 | 180 | 148 |
| Proliferative diabetic retinopathy | 571 | 85 | 176 |
| Wet age-related macular degeneration | 46 | 86 | 36 |
| Pathologic myopia | 96 | 49 | 140 |

**Table S5** Distribution of tumor categories in the internal and external UWF tumor cohorts. The external cohort contained no nevus cases.

| Tumor category | Optos-UWF-Tumor | External-Optos-Tumor |
|---|---|---|
| Health | 413 | 1354 |
| Melanoma (Mel) | 1026 | 85 |
| Nevus (Nev) | 638 | 0 |

# 6 Effect of restoring the original three-channel RGB input

To evaluate whether adapting the input projection layer to accommodate two-channel UWF retinal images disadvantaged foundation models originally pretrained on three-channel RGB images, we additionally evaluated all baseline models using their original three-channel RGB inputs without modifying the input projection layer, (Table S8). For this control experiment, the original exported UWF images were used directly, in which the third RGB channel contained zero-valued pixels and therefore carried no additional retinal information. Restoring the original RGB input substantially improved RETFound under both full fine-tuning (+7.26 mAUC) and frozen-feature evaluation (+9.09 mAUC), whereas only marginal changes were observed for MAE-In1K (+1.40 mAUC and +0.02 mAUC, respectively). DINOV3 showed no consistent benefit from RGB inputs. Importantly, despite these improvements, UWF-FM remained the best-performing model under both full fine-tuning and frozen-feature evaluation, indicating that the performance advantage of UWF-FM cannot be attributed solely to the two-channel input adaptation strategy adopted for baseline models.

**Table S6** Distribution of diabetic retinopathy (DR) severity grading categories across the internal development cohort (Optos-UWF-DR) and independent external validation cohorts from Algeria and Shanghai. The internal DR grading dataset comprised images collected from both Xiamen and Zhejiang clinical centers.

| DR Grade | Optos-UWF-DR | External-Optos-DR | External-Optos-DR1 |
|---|---|---|---|
| No DR | 1944 | 1214 | 74 |
| Mild NPDR | 636 | 347 | 73 |
| Moderate NPDR | 707 | 115 | 71 |
| Severe NPDR | 736 | 70 | 28 |
| PDR | 596 | 15 | 4 |

**Table S7** Distribution of diabetic retinopathy quality control (QC) categories in the Optos-UWF-QC dataset. Images may be assigned to multiple QC categories.

| QC category | Optos-UWF-QC |
|---|---|
| Diabetic retinopathy | 3724 |
| Refractive media opacity | 462 |
| Laser scars | 423 |
| Preretinal hemorrhage | 404 |
| Vitreous hemorrhage | 592 |
| Fibrovascular proliferation | 494 |
| Ungradable image | 87 |
| Other retinal diseases | 456 |

# 7 Effect of input resolution

To evaluate whether the performance advantage of UWF-FM depended on the choice of input resolution, we repeated the retinal finding characterization experiments using input resolutions of (384×384) and (512×512), while maintaining identical downstream training protocols and evaluation settings (Supplementary Fig. S9).

Increasing the input resolution produced only modest changes in performance for both UWF-FM and the strongest ImageNet-pretrained baseline (MAE-In1K). Under full fine-tuning, performance improved slightly when increasing the resolution from (224×224) to (384×384), whereas further increasing the resolution to (512×512) provided little additional benefit. Similar trends were observed under the frozen-feature (Probe) setting, where higher input resolutions resulted in only minor performance variations.

Importantly, UWF-FM consistently outperformed MAE-In1K across all evaluated input resolutions under both full fine-tuning and frozen-feature evaluation. These results indicate that the superiority of UWF-FM is robust to the choice of input

**Table S8** Control experiment evaluating the effect of restoring the original three-channel RGB input for baseline foundation models. Baseline models were evaluated using either the two-channel UWF input adopted throughout this study or their original three-channel RGB/JPEG input without modifying the input projection layer. Results are reported on the Optos-UWF-Lesion using identical downstream training protocols.

| Model | **FT** | | **Probe** | |
|---|---|---|---|---|
| | 2-channel | RGB | 2-channel | RGB |
| DINOV3 | 74.68 ± 2.77 | 73.89 ± 11.59 | 56.88 ± 4.72 | 63.36 ± 2.39 |
| RETFound | 74.70 ± 0.86 | 81.96 ± 0.10 | 63.82 ± 0.68 | 72.91 ± 0.49 |
| MAE-In1K | 82.48 ± 2.72 | 83.88 ± 0.63 | 69.94 ± 0.42 | 69.96 ± 0.43 |
| **UWF-FM** | **87.77 ± 0.33** | – | **83.39 ± 0.47** | – |

resolution and cannot be attributed to the use of a specific image resolution during downstream training.

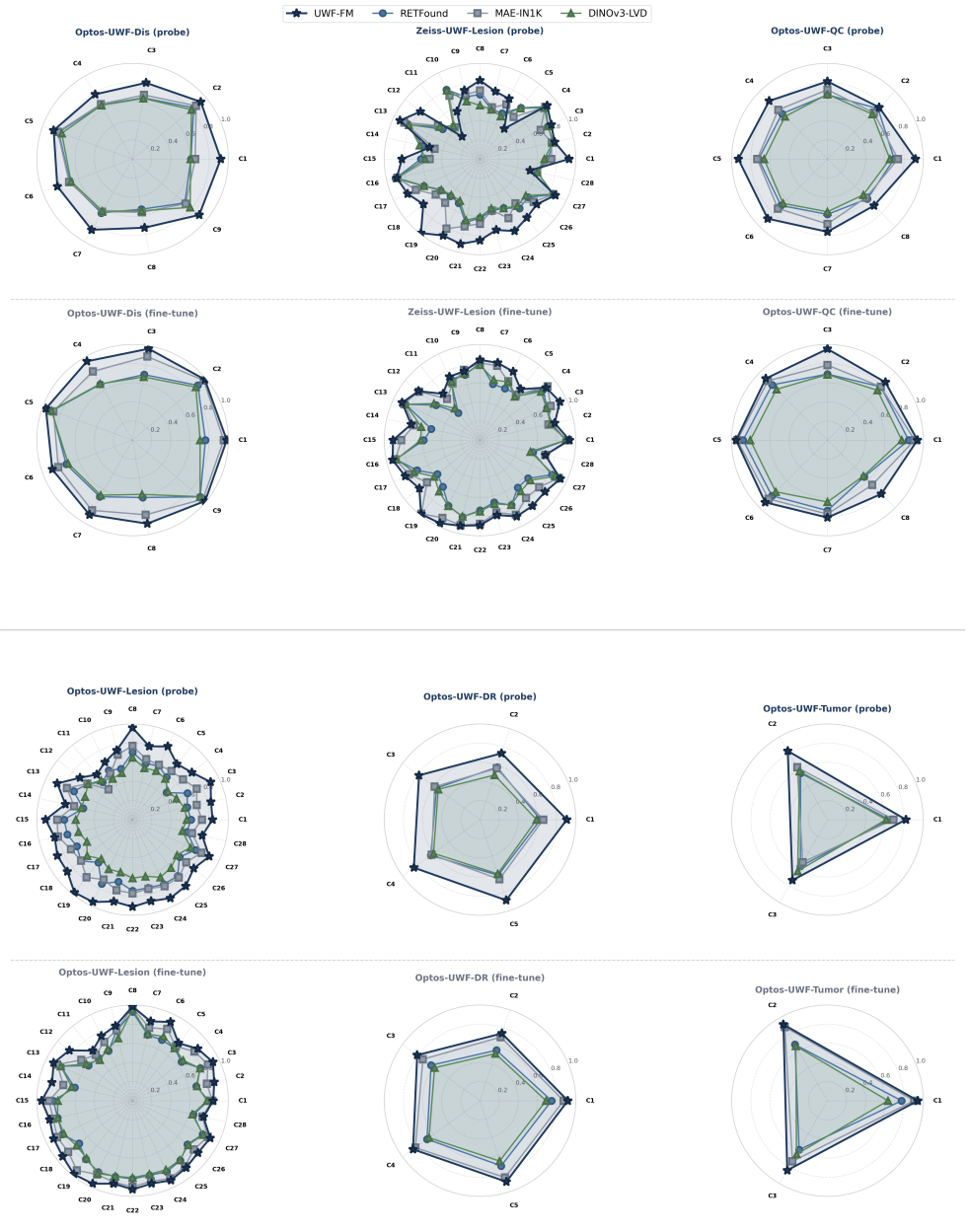


**Fig. S1 Category-level performance comparison across clinically structured retinal assessment settings.** Each axis represents an individual disease, DR assessment, or retinal finding category. Lines indicate mean AUCs of different pretrained models under (a) full fine-tuning and (b) frozen-feature evaluation. UWF-FM consistently achieves superior performance across the majority of categories.

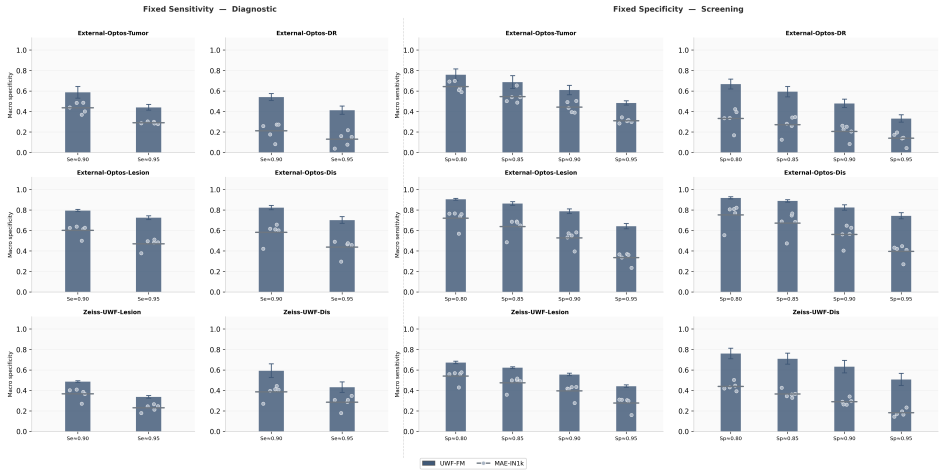


**Fig. S2 Performance across clinically relevant operating thresholds.** Sensitivity is shown at fixed specificity thresholds (80%, 85%, 90%, and 95%) across internal and external UWF retinal assessment settings. Points indicate mean sensitivity and error bars represent standard deviations across five-fold cross-validation.

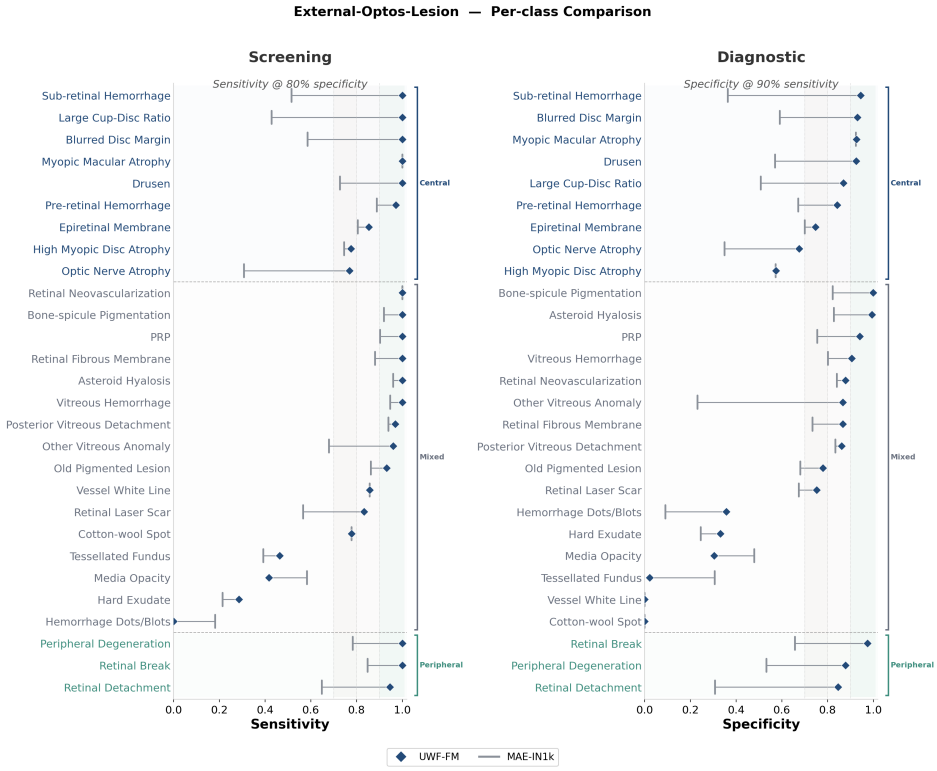


**Fig. S3 Category-level operating characteristics on the External-Optos-Lesion cohort.** The left panel shows sensitivity at 80% specificity (screening-oriented setting), and the right panel shows specificity at 90% sensitivity (diagnostic-support setting). Each row represents an individual retinal finding category. UWF-FM is shown as diamonds and MAE-In1K as horizontal ticks, with lines connecting paired category-level estimates. Categories are arranged according to their predominant retinal distribution: Central (Predominantly posterior-pole), Mixed (Retina-wide), and Perpheral (Predominantly peripheral).

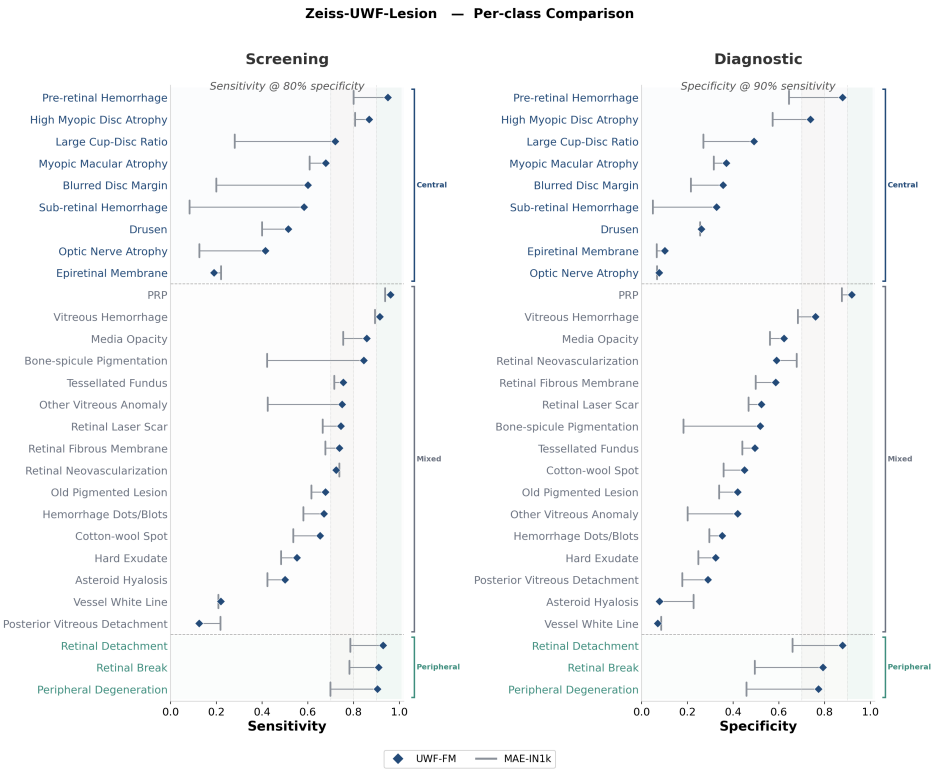


**Fig. S4 Category-level operating characteristics on the Zeiss-UWF-Lesion cohort.** The left panel shows sensitivity at 80% specificity (screening-oriented setting), and the right panel shows specificity at 90% sensitivity (diagnostic-support setting). Each row represents an individual retinal finding category. UWF-FM is shown as diamonds and MAE-In1K as horizontal ticks, with lines connecting paired category-level estimates. Categories are arranged according to their predominant retinal distribution: : Central (Predominantly posterior-pole), Mixed (Retina-wide), and Peripheral (Predominantly peripheral).

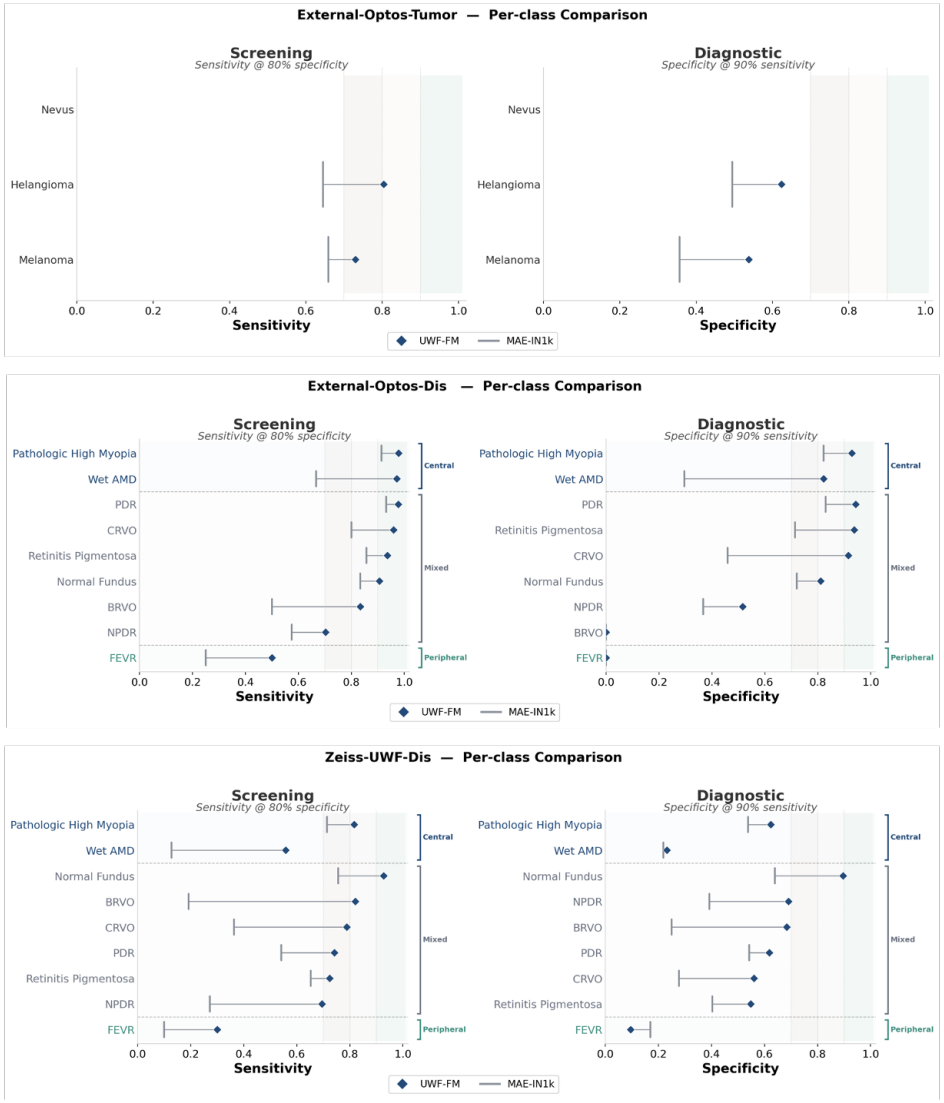


**Fig. S5 Category-level operating characteristics for disease diagnosis.** The left panel shows sensitivity at 80% specificity and the right panel shows specificity at 90% sensitivity. Each row represents an individual disease category. UWF-FM is shown as diamonds and MAE-In1K as horizontal ticks, with lines connecting paired category-level estimates.

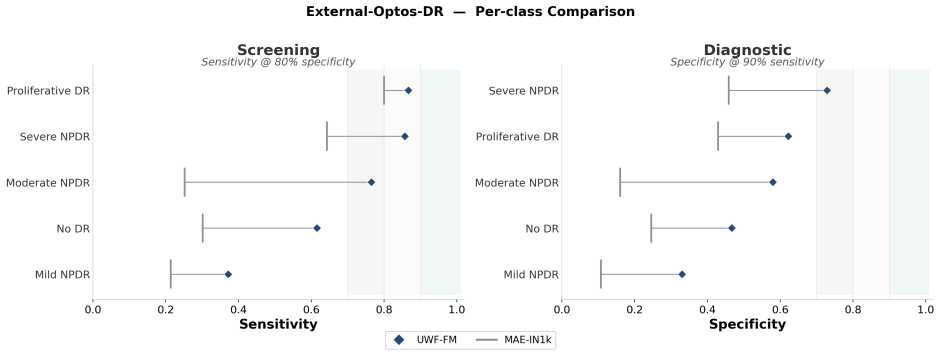


**Fig. S6 Category-level operating characteristics for DR assessment.** The left panel shows sensitivity at 80% specificity and the right panel shows specificity at 90% sensitivity. Each row represents an individual DR severity or image-quality category. UWF-FM is shown as diamonds and MAE-In1K as horizontal ticks, with lines connecting paired category-level estimates.

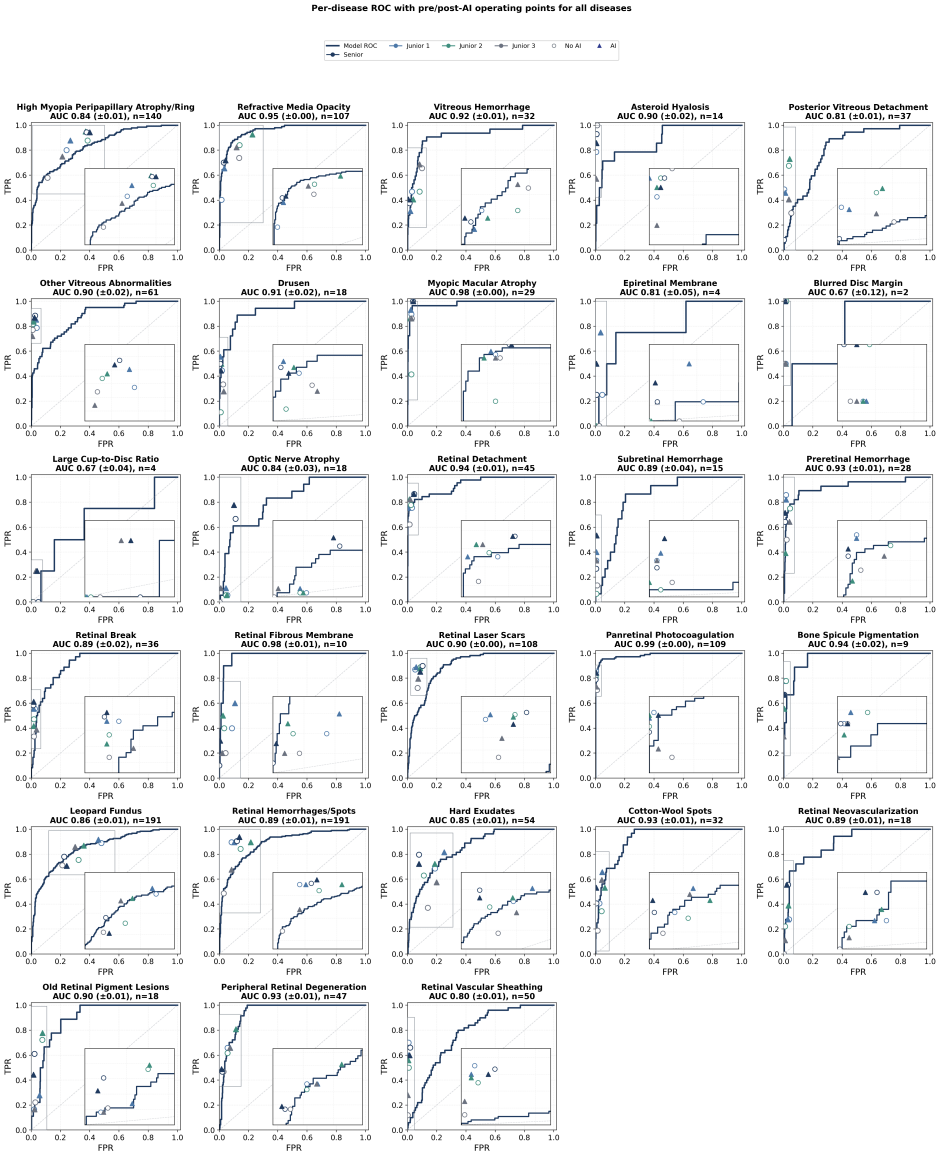


**Fig. S7 Retinal finding-level ROC curves with clinician operating points before and after AI assistance.** Receiver operating characteristic (ROC) curves for individual retinal findings, with clinician operating points shown for unaided (pre-AI) and AI-assisted (post-AI) conditions. Each panel corresponds to one retinal finding category. AI assistance results in variable shifts in operating points across retinal findings, reflecting heterogeneity in task difficulty and model–clinician interaction.

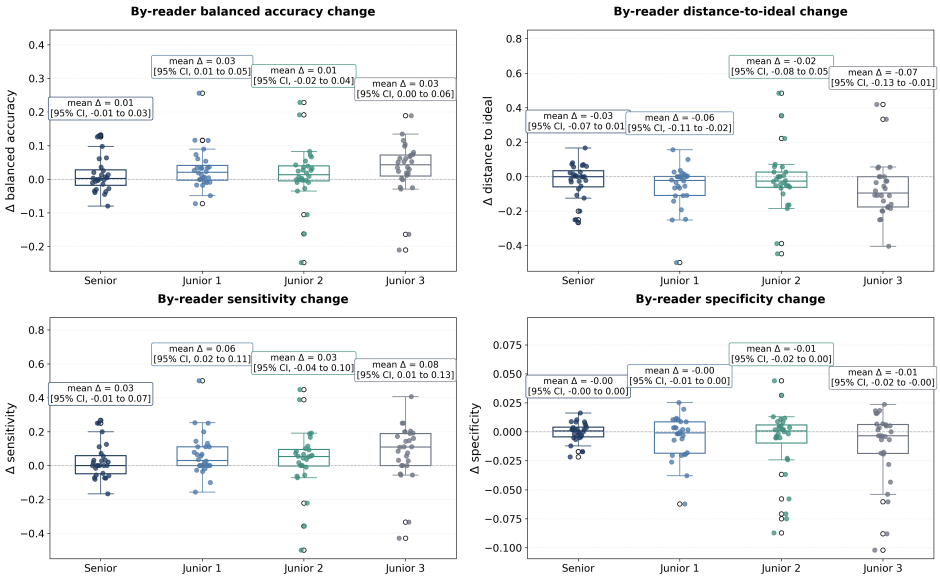


**Fig. S8 By-reader summary of AI-associated performance changes.** Distribution of retinal finding-level performance changes for each clinician, including balanced accuracy, distance to the ideal operating point, sensitivity, and specificity. Each point represents the performance change associated with one retinal finding category for a given reader. AI assistance leads to heterogeneous effects across clinicians and within individual readers, with variability in both the magnitude and direction of performance changes.

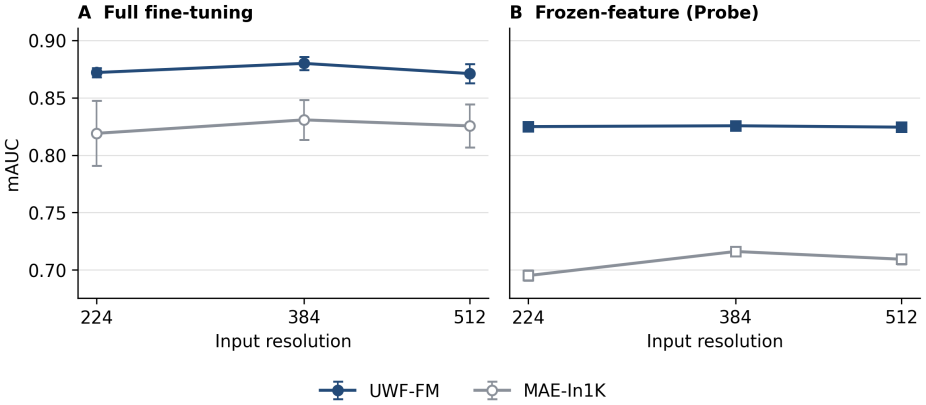


**Fig. S9 Sensitivity of model performance to input resolution.** Performance of UWF-FM and the strongest ImageNet-pretrained baseline (MAE-In1K) on retinal finding characterization using input resolutions of (224×224), (384×384), and (512×512). a, Full fine-tuning. b, Frozen-feature (Probe). Error bars represent the standard deviation across repeated runs. UWF-FM consistently outperformed MAE-In1K across all evaluated resolutions, indicating that the observed performance advantage was robust to the choice of input resolution.