Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predicting Visual Acuity from Fundus Photographs Using a Domain-Pretrained Vision Transformer: Anatomical Alignment of Attention Patterns

Jin Hyun Kim
Gyeongsang National University
Yong Seop Han
medcabin@naver.com

Hanmaeum Hospital
Kuk Jin Jang
Hongik University
Seoung Jin Lee
Gyeongsang National University
Hyonyoung choi
University of Pennsylvania
Insup Lee
University of Pennsylvania

Research Article

Keywords: visual acuity, fundus photography, vision transformer, RETFound, attention rollout, ophthalmic AI

Posted Date: July 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10085109/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Predicting Visual Acuity from Fundus Photographs Using a Domain-Pretrained Vision Transformer: Anatomical Alignment of Attention Patterns

Jin Hyun Kim[1], Yong Seop Han[2*], Kuk Jin Jang[3], Seoung Jin Lee[1], Hyonyoung choi[3], Insup Lee[3]

[1]AI Information Engineering, Gyeongsang National Univ., Jinju, S.Korea.
[2]Department of Ophthalmology, Changwon Hanmaeum Hospital., Changwon, S.Korea.
[3*]Department of Computer Engineering, Hongik Univ.,, Seoul, S.Korea.
[3]PRECISE Center and Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA.

*Corresponding author(s). E-mail(s): medcabin@naver.com;
Contributing authors: jin.kim@gnu.ac.kr; jangkj@hongik.ac.kr; insight@gnu.ac.kr; hyonchoi@seas.upenn.edu; lee@seas.upenn.edu;

Abstract

Purpose. To evaluate a domain-pretrained Vision Transformer for four-class ordinal visual acuity (VA) prediction from single fundus photographs and to characterize the anatomical organization of learned attention patterns.

Methods. We retrospectively collected 21,602 fundus photographs from 3,654 patients with concurrent best-corrected VA measurements. Images were classified into four ordinal VA classes (C0: 0.0; C1: 0.1-–0.2; C2: 0.3—0.7; C3: 0.8—1.0). RETFound (ViT-L/16, 307M parameters), pretrained on 1.6 million unlabeled fundus images, was fine-tuned under a progressive unfreezing schedule; the primary architecture comparison used OrdinalFocalLoss, while additional loss ablations compared class-weighted cross-entropy, focal cross-entropy, and EMD. Patient-level splits ensured a fixed held-out test set (564 patients, 3,236 images). Image-level and exam-level metrics were reported, where exam_id was defined as patient+date+eye. Attention Rollout was computed across all 24 transformer blocks and quantified using anatomically motivated masks (retinal vessels via Frangi filter; macula via circular mask at image center). Robustness analyses included 180-combination mask sensitivity, negative controls (spatial

shuffling and random-weight ViT), cross-method comparison (GradCAM++), discard-ratio sensitivity, and macular-center offset sensitivity.
Results. RETFound achieved 55.6% accuracy (95% CI: 53.9–57.2%), MAE 0.493 (95% CI: 0.473–0.514), weighted F1 0.544 (95% CI: 0.526–0.561), and quadratic weighted kappa (QWK) 0.612 (95% CI: 0.587–0.635; "substantial agreement" by Landis & Koch criteria), outperforming DINOv2 (ViT-B/14-reg4) and EfficientNet-B4 (McNemar $p < \mathbf{0.0001}$ for both). Of all RETFound errors, 89.6% were confined to adjacent VA classes ($|\mathbf{pred} - \mathbf{true}| = \mathbf{1}$). In loss ablation on the RETFound backbone, class-weighted cross-entropy yielded the best weighted F1 (0.547) and QWK (0.617), whereas OrdinalFocalLoss achieved the best top-1 accuracy (55.6%) and the lowest MAE (0.493), indicating a trade-off between agreement-oriented and ordinal-distance-oriented objectives. Vessel attention enrichment decreased monotonically with VA severity (C0: 1.748 → C3: 0.702; Kruskal-Wallis $p=4.88\times\mathbf{10^{-11}}$; $\boldsymbol{\eta^2}=0.254$), while macular enrichment increased (C0: 0.706 → C3: 1.165; $p=4.07\times\mathbf{10^{-8}}$; $\boldsymbol{\eta^2}=0.181$). Both patterns were stable across all 180 mask parameter combinations. Correctly classified samples showed markedly stronger log-ADR–VA correlation (Spearman $\boldsymbol{\rho}=0.644$, 95% CI [0.553, 0.720]) than misclassified samples ($\boldsymbol{\rho}=0.198$, 95% CI [0.063, 0.326]).
Conclusions. RETFound fine-tuned for VA classification develops anatomically organized attention that mirrors established ophthalmic anatomy, with vessel attention dominating in low-VA and macular attention in normal-VA eyes. The alignment between anatomical attention patterns and classification correctness suggests a potential role for attention-based interpretability in flagging uncertain predictions.

Keywords: visual acuity, fundus photography, vision transformer, RETFound, attention rollout, ophthalmic AI

1 Introduction

Visual acuity (VA) is the primary clinical measure of functional vision and a central endpoint in ophthalmic diagnosis, treatment monitoring, and disability assessment [1]. Accurate VA measurement conventionally requires patient cooperation and standardised chart examination—conditions that may not always be feasible in population-level screening, telemedicine deployments, or resource-limited settings. Fundus photography, by contrast, is a non-invasive, widely deployed imaging modality that captures the structural state of the retina with high reproducibility across devices and operators [2]. Whether the rich anatomical information encoded in a single fundus photograph is sufficient to infer VA class—and whether a deep learning model can learn to attend to clinically relevant structures in doing so—remains an open and practically consequential question.

Prior work has explored automated VA prediction from retinal images using convolutional neural networks (CNNs), typically framing the task as binary classification (impaired vs. normal) or as continuous regression [3, 4]. These approaches, while demonstrating technical feasibility, do not reflect the ordinal, multi-class structure of

clinical VA grading—a structure that carries direct prognostic and therapeutic implications. Moreover, most existing models treat prediction as a black box, providing no insight into which retinal regions drive the decision. In clinical ophthalmology, the spatial origin of a diagnostic signal matters: peripheral vascular pathology, macular structural integrity, and optic disc changes each correspond to distinct disease mechanisms and management strategies. A model that learns to weight these structures appropriately would not only be more interpretable, but would also provide a stronger guarantee that it is learning clinically meaningful features rather than spurious image statistics.

Our own group has previously addressed this task in two studies. Kim et al. [5] proposed an ensemble method combining support vector machines and two CNN architectures to classify fundus images into four VA levels, reporting an average accuracy of 82.4% on a held-out set of 4,000 images from a single-institution database. Subsequently, Lee et al. [6] extended this line of work with a larger-scale CNN framework trained on 53,318 fundus photographs from the same institution, achieving 94.37% top-3 tolerance accuracy. While these results were encouraging, a critical re-examination of both studies revealed a fundamental methodological flaw: the train–test splits were performed at the image level rather than the patient level, meaning that multiple images from the same patient could appear in both training and test sets. This patient-level data leakage inflates reported performance figures, as the model may effectively memorise patient-specific retinal characteristics rather than learning generalisable representations of VA-related anatomy. The present study was designed directly to address this limitation, enforcing strict patient-level splits and a fixed held-out test set to provide an unbiased estimate of true generalisation performance.

Foundation models pretrained on large, domain-specific datasets have recently demonstrated exceptional transfer performance across medical imaging tasks [7]. RETFound—a Vision Transformer (ViT-L/16) pretrained on 1.6 million unlabelled fundus photographs via masked autoencoding (MAE) [8]—is a particularly well-suited backbone for this problem: its ophthalmology-specific pretraining encodes low-level retinal texture, vascular morphology, and structural landmarks that general-purpose models such as DINOv2 [9] cannot capture from natural-image datasets alone. However, whether domain-specific pretraining translates into anatomically organised internal representations—and whether those representations align with established clinical knowledge about the retinal determinants of VA—has not been systematically investigated.

Attention Rollout [10] offers a principled method for tracing spatial attention flow across all transformer layers, aggregating the cumulative influence of each image patch on the classification token. Crucially, unlike gradient-based attribution methods such as GradCAM++ [11], Rollout characterises where the model looks rather than what drives the classification decision—a distinction with direct implications for clinical interpretability. Yet quantitative validation of Rollout maps against anatomically defined regions, and analysis of whether the resulting attention patterns are functionally linked to prediction correctness at the individual sample level, remain largely absent from the ophthalmic AI literature.

In this study, we address these gaps with three principal contributions. First, we fine-tune RETFound for four-class ordinal VA prediction from single fundus photographs under a rigorous patient-level cross-validation protocol, and demonstrate statistically significant superiority over a general-purpose ViT (DINOv2, ViT-B/14-reg4) and a convolutional baseline (EfficientNet-B4), with 89.6% of errors confined to adjacent VA classes (quadratic weighted $\kappa = 0.612$). Second, we characterise the anatomical organisation of learned attention using quantitative enrichment analysis across 180 mask-parameter combinations, two negative controls, and a cross-method comparison with GradCAM++, establishing that vessel attention dominates in low-VA eyes and macular attention dominates in normal-VA eyes—a pattern that mirrors established ophthalmic anatomy. Third, we introduce individual-level attention analysis via a log Attention Dominance Ratio (log-ADR), showing that the alignment between anatomical attention and correct classification is 3.3-fold stronger in correctly classified samples (Spearman $\rho = 0.644$) than in misclassified ones ($\rho = 0.198$; non-overlapping 95% CIs), suggesting a potential role for attention-based interpretability in flagging uncertain predictions at the point of care.

2 Methods

2.1 Dataset and Study Population

We retrospectively collected fundus photographs from patients who underwent ophthalmic examination with concurrent best-corrected visual acuity (BCVA) measurement. BCVA was recorded as a decimal fraction (0.0–1.0) using standardized Landolt C or Snellen chart examination; values were converted to the nearest decimal representation and assigned to four ordinal classes based on clinical severity:

Class	VA Range (decimal)	Clinical Category
C0	0.0	Total/functional blindness
C1	0.1–0.2	Severe visual impairment (near legal blindness level)
C2	0.3–0.7	Moderate-to-mild impairment
C3	0.8–1.0	Normal vision

The final dataset comprised **21,602 images from 3,654 patients** (mean ± SD: 5.9 ± 4.2 images per patient; range 1–38). Both eyes were included where available; the mean number of images per eye was 3.1 ± 2.4. To prevent data leakage, all splits were performed at the **patient level** using a unique patient identifier extracted from each filename. The held-out test set (564 patients, 3,236 images) was fixed prior to any model development. The remaining 3,090 patients (18,366 images) were divided using 2-fold cross-validation for training and internal validation. The test set class distribution reflected the natural clinical prevalence: C0 9.6% (n=310 images), C1 16.2% (n=525),C2 46.6% (n=1,507), C3 27.6% (n=894). For post hoc aggregation analysis, an **exam_id** was defined as patient_id + acquisition_date + eye (OD/OS); all 3,236 held-out images corresponded to unique exam_ids, so exam-level aggregation was numerically identical to image-level evaluation.

[**TO VERIFY**] VA measurement device and chart type should be confirmed from clinical record metadata for the final submission.

2.2 Image Preprocessing

All fundus photographs were preprocessed with Contrast Limited Adaptive Histogram Equalization (CLAHE; clip limit = 2.0, tile grid = 8×8) applied to the luminance channel in LAB color space to standardize retinal contrast across imaging devices. Images were resized to 224×224 pixels and normalized using channel-wise mean = [0.452, 0.278, 0.166] and std = [0.131, 0.100, 0.077], computed from the **training subset only** (3,090 patients, 18,366 images) to prevent any statistical leakage from the held-out test set.

Training-time augmentation included random horizontal/vertical flipping (p=0.5 each), random rotation (±30°), random affine transforms (translation ±5%, scale 0.9–1.1), color jitter (brightness/contrast/saturation=0.3, hue=0.1), Gaussian blur (σ=0.1–1.5), and random erasing (p=0.2, area 2–15%). At inference, test-time augmentation (TTA) was applied as the mean of four predictions (original, horizontal flip, vertical flip, both flips combined) for all evaluated models.

2.3 Model Architecture

We used RETFound (Retinal Foundation Model), as shown in Fig 1, a Vision Transformer Large/16 (ViT-L/16; 24 transformer blocks, 1024-dimensional embeddings, 16×16 pixel patch size, 307M parameters) pretrained on 1.6 million unlabeled fundus photographs using masked autoencoding (MAE). A four-layer classification head was appended:

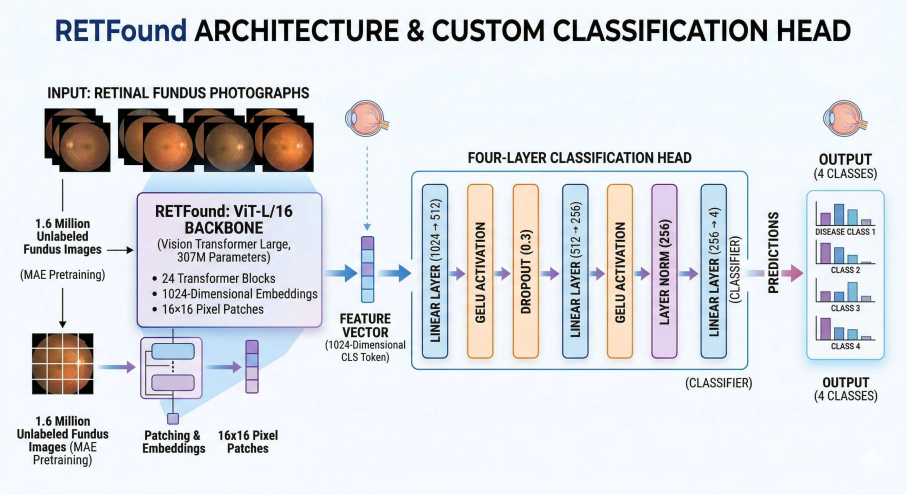


Fig. 1: RetFound Architecture and Custom Classification Head

To enable interpretable attention extraction, fused attention was disabled (`fused_attn = False`) in all transformer blocks.

Comparison models.

For baseline comparison, (1) DINOv2 (ViT-B/14-reg4; 86M parameters; Vision Transformer Base with 14×14 pixel patches and 4 register tokens; native input resolution 518×518; pretrained on the LVD-142M curated image dataset using self-distillation with no labels [DINO v2]) and (2) EfficientNet-B4 (19M parameters; convolutional backbone) were fine-tuned under identical training conditions (same optimizer, scheduler, loss function, augmentation, and TTA) as described in Section 2.4. Training conditions for all models are summarized in Supplementary Table S1.

2.4 Training Procedure

Loss function. We employed an OrdinalFocalLoss combining focal cross-entropy with an ordinal regression penalty:

$$\mathcal{L} = -\alpha_t(1 - p_t)^{\gamma} \log p_t \; + \; \lambda_{\text{ord}} \cdot \text{MSE}(\mathbb{E}[\hat{y}],\, y)$$

where $\gamma = 2.0$ (down-weighting the majority class C2 at 46.6%), λ_ord = 1.0, and α_t = inverse class frequency. The focal term down-weights easy examples and concentrates learning on hard cases; the ordinal MSE term penalizes large-magnitude ordinal errors by training on the expected class value under the predicted softmax distribution. This loss was applied uniformly to **the main architecture comparison models** (RETFound, DINOv2, EfficientNet-B4). Its relative contribution versus class-weighted cross-entropy, focal cross-entropy, and Earth Mover's Distance (EMD), a distance-based loss that accounts for ordinal misclassification severity, is reported in the loss ablation experiments (Section 3.1, Table 4).

Optimization. AdamW optimizer (weight decay = 1e-4) with ReduceLROnPlateau scheduler (factor = 0.5, patience = 7 epochs). Gradient accumulation over 2 steps yielded an effective batch size of 16 (mini-batch = 8).

Progressive unfreezing. To preserve pretrained fundus representations while adapting to the VA task:

Phase	Epochs	Trainable Parameters	Head LR	Backbone LR
0	0–7	Head only	1e-3	—
1	8–27	Head + last 25% blocks	1e-3	2e-5
2	28–59	Head + last 50% blocks	1e-3	1e-5

Early stopping with patience = 20 epochs; best model selected by validation weighted-F1. Progressive unfreezing was applied to RETFound only; DINOv2 and EfficientNet-B4 used standard full-model fine-tuning from epoch 0 (see Supplementary Table **??**).

2.5 Attention Rollout

Transformer-based models distribute information across image patches through self-attention, but raw attention weights at a single layer are not reliable indicators of

input relevance: as depth increases, information from different tokens becomes progressively mixed, making any single layer's attention map an incomplete picture of what the model ultimately relies upon [10]. Attention Rollout addresses this by propagating attention matrices sequentially from the first to the last transformer block via recursive matrix multiplication, thereby accounting for the residual connections and the cumulative mixing of information across all layers. The result is a single spatial map—assigned to the classification (`CLS`) token—that approximates the aggregate influence of each input patch on the final representation. We adopt Attention Rollout here to answer the spatial question: which retinal regions does the fine-tuned model attend to when predicting each VA class?

To compute Attention Rollout, we processed all 24 transformer blocks sequentially. At each block ℓ, the multi-head attention matrix $\mathbf{A}^{(\ell)} \in \mathbb{R}^{N \times N}$ (where N is the number of tokens including the `CLS` token) was first averaged across all attention heads. The averaged matrix was then thresholded by setting the lowest $(1 - r_{\text{discard}})$ fraction of weights to zero (discard ratio $r_{\text{discard}} = 0.9$, retaining only the top 10% of attention weights to focus the rollout on the most informative attention pathways; the sensitivity of all enrichment findings to this parameter is systematically evaluated across $r_{\text{discard}} \in \{0.7, 0.8, 0.9, 0.95\}$ in Section 2.9.4). A residual connection was incorporated by adding the identity matrix, and the result was normalised row-wise:

$$\tilde{\mathbf{A}}^{(\ell)} = \text{normalise_rows}\left(\hat{\mathbf{A}}^{(\ell)} + \mathbf{I}\right), \tag{1}$$

where $\hat{\mathbf{A}}^{(\ell)}$ denotes the thresholded, head-averaged attention matrix at block ℓ. The cumulative rollout map was computed by sequential matrix multiplication across all blocks:

$$\mathbf{R} = \tilde{\mathbf{A}}^{(24)} \cdot \tilde{\mathbf{A}}^{(23)} \cdots \tilde{\mathbf{A}}^{(1)}. \tag{2}$$

The row of $\mathbf{R}$ corresponding to the `CLS` token was extracted, reshaped into a 14×14 patch grid (consistent with the ViT-L/16 patch size), and upsampled to 224×224 using bilinear interpolation to produce the final spatial attention map. All rollout computations were performed under `torch.no_grad()` to suppress gradient tracking and reduce memory overhead.

To quantify spatial properties of the resulting rollout maps, we computed two complementary metrics per sample.

Center Proximity Ratio (CPR).

CPR is defined as the fraction of total attention weight located within a circular region centred at the image centre with radius equal to half the image half-width ($r_{\text{CPR}} = 56\,\text{px}$ for a 224×224 image):

$$\text{CPR} = \frac{\sum_{i \in \mathcal{D}_{56}} a_i}{\sum_i a_i}, \tag{3}$$

where $\mathcal{D}_{56}$ denotes the set of pixels within 56 px of the image centre and a_i is the rollout attention weight at pixel i after ℓ_1-normalisation of the upsampled map. A higher CPR indicates attention concentrated toward the foveal region, while a lower CPR reflects peripheral distribution.

Spatial Entropy.

Spatial Entropy measures the degree of diffuseness of the attention distribution via the Shannon entropy of the normalised map:

$$H = -\sum_i a_i \log(a_i + \varepsilon), \qquad \varepsilon = 10^{-9}, \tag{4}$$

where a_i is as defined above and ε prevents numerical instability at near-zero attention weights. Higher entropy indicates broadly distributed attention across multiple retinal regions; lower entropy indicates focal concentration. Together, CPR and Spatial Entropy provide complementary perspectives on whether attention is both centralised (CPR) and concentrated (entropy) within the image.

2.6 Anatomical Mask Extraction

Three binary anatomical masks were delineated on each 224×224 fundus image to spatially characterise the model's attention relative to clinically meaningful retinal structures.

Retinal vessels.

A Frangi vesselness filter [12] (scale range $\sigma = 1$–4 px; `black_ridges=False`) was applied to the grayscale image to enhance tubular retinal structures. Pixels whose vesselness response exceeded a threshold of 0.01 were classified as vessel pixels (mean area fraction $\approx 15.8\%$ of image pixels). The sensitivity of findings to this threshold is evaluated in Section 2.9.1.

Macula.

The macular region was approximated by a fixed circular mask centred at the geometric centre of the image (radius = 33 px; area fraction $\approx 8.8\%$). This fixed-centre approximation is valid because the macula is reliably centred in standardised non-mydriatic and mydriatic fundus photographs. Robustness to foveal displacement is evaluated in Section 2.9.5.

Optic disc.

The optic disc was identified as the brightest region (top 5% luminance) within a 25 px radius search zone centred at the image centre. As illustrated in Fig. 2, this approach correctly localises the disc in standard fundus photographs, where the optic disc appears as the dominant bright structure in the nasal retinal field. The optic disc mask is used for the dominant structure analysis only (Section 2.8) and is not included in the primary enrichment interpretation.

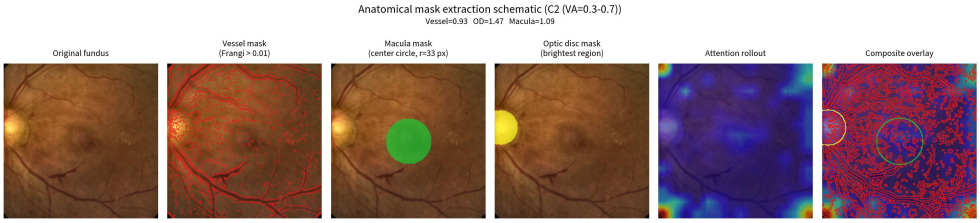


Fig. 2: Anatomical Mask Extraction Schematic

A representative schematic of the extraction pipeline is shown in Fig. 2, illustrating the six panels from left to right: the CLAHE-preprocessed fundus image; the Frangi vessel mask (vessel pixels in red); the macular mask (filled green circle, $r = 33\,\text{px}$); the optic disc estimate (filled yellow circle); the Attention Rollout heatmap (cool-to-warm scale); and the composite overlay combining the rollout map with all three anatomical mask boundaries. In this C2 example, the attention rollout is relatively diffuse with no single structure dominating, consistent with the intermediate VA class. The optic disc mask (panel 4) is visibly displaced from the true disc margin in this image, confirming that the brightness heuristic provides only a coarse approximation and justifying its exclusion from the primary enrichment analysis.

2.7 Attention Enrichment

For each anatomical region, the enrichment ratio was computed as:

$$E_{\text{region}} = \frac{\sum_{i \in \text{mask}} A_i \;/\; \sum_{i} A_i}{\text{mask area/total pixels}}, \tag{5}$$

where A_i is the rollout attention weight at pixel i after ℓ_1-normalisation of the upsampled attention map. A value of $E = 1.0$ indicates attention proportional to the mask area (random baseline); $E > 1.0$ indicates relative concentration of attention within the region; and $E < 1.0$ indicates depletion. Raw attention ratios (the numerator of Eq. 5 before area normalisation) are provided in Supplementary Figure 1 (Fig **??**) for reference.

2.8 Individual-Level Analysis

To move beyond group-level means, each sample was characterised using a log Attention Dominance Ratio (log-ADR), defined as:

$$\text{log-ADR} = \log(\text{macula_enrichment} + \varepsilon) - \log(\text{vessel_enrichment} + \varepsilon)\,, \qquad \varepsilon = 0.01, \tag{6}$$

where ε prevents numerical instability at near-zero enrichment values. A positive log-ADR indicates macular attention dominance over vessels; a negative value indicates vessel dominance.

Each sample was further assigned to a dominant structure category (vessel, optic disc, or macula) if the enrichment of that region exceeded those of both other regions

by at least 0.25 and its absolute enrichment was at least 1.15. Samples were labelled *mixed* when the two highest enrichment values were within 0.25 of each other, or *diffuse* when all three enrichment values fell below 1.15.

To assess the linear discriminability of the enrichment features for VA class prediction, a logistic regression probe was trained on the two-dimensional feature vector $[\log(\text{vessel_enrichment} + \varepsilon),\ \log(\text{macula_enrichment} + \varepsilon)]$ using L2 regularisation ($C = 1.0$) with 5-fold stratified cross-validation.

2.9 Validation Experiments

2.9.1 val_01 — Parameter sensitivity

To assess the robustness of enrichment findings to mask parameterisation, the enrichment analysis was repeated across 180 parameter combinations: 4 vessel thresholds (0.005, 0.01, 0.02, 0.05) × 5 macular radii (20, 27, 33, 40, 45 px) × 3 disc radii × 3 disc brightness percentiles. Rollout maps were computed once per sample and reused across all combinations to isolate mask sensitivity from attention computation. Two retention metrics were evaluated across all 180 combinations: (1) *direction retention rate* — the fraction of combinations in which vessel enrichment was higher in C0 than C3 and macular enrichment was higher in C3 than C0 simultaneously; and (2) *KW significance retention* — the fraction of combinations yielding a Kruskal–Wallis $p < 0.05$ for both structures.

2.9.2 val_03 — Negative controls

Two negative controls were applied to verify that the observed enrichment patterns arise from learned spatial attention rather than image statistics or random model weights.

- *Spatial control (Shuffle):* attention weights were pixel-wise permuted (seed = 42), destroying all spatial structure while preserving the marginal weight distribution. Under this control, enrichment ratios should converge to 1.0 for all regions and VA classes.
- *Learning control (Random weight):* enrichment was computed from a ViT-L/16 with random weight initialisation (no pretrained checkpoint), matched in architecture to the fine-tuned model. This control dissociates enrichment patterns attributable to learned features from those arising from architectural inductive biases alone.

Both controls were evaluated on a balanced subsample of 100 images per class (seed = 42) drawn from the held-out test set.

2.9.3 val_02 — Cross-method validation

To examine whether the spatial patterns observed in Attention Rollout are corroborated by a gradient-based attribution method, GradCAM++ [11] was applied to the final normalisation layer (`model.blocks[-1].norm1`) using a ViT-specific reshape

transform (14×14 patch grid; CLS token discarded). The ground-truth class was used as the attribution target throughout, ensuring class-consistent gradient computation and avoiding confounding by model misclassification (which is particularly relevant given the low C0 recall of 43%). Enrichment was computed on the same balanced subsample of 400 images (100 per class; seed = 42) with identical anatomical masks as used in the primary analysis.

2.9.4 val_04 — Discard ratio sensitivity

To evaluate the sensitivity of enrichment findings to the Attention Rollout discard ratio, this parameter was varied across four values ($r_{\text{discard}} \in \{0.7, 0.8, 0.9, 0.95\}$), corresponding to retaining the top 30%, 20%, 10%, and 5% of attention weights, respectively. All other parameters were held constant (vessel threshold = 0.01; macular radius = 33 px; balanced 100 images per class; seed = 42). Rollout maps were recomputed independently for each discard ratio value and enrichment statistics recalculated.

2.9.5 val_05 — Macular centre offset sensitivity

To evaluate robustness to potential foveal displacement from the image centre, the macular mask centre was varied across a 5×5 grid of offset positions (Δx, $\Delta y \in \{-20, -10, 0, +10, +20\}$ px; 25 positions in total), with the mask radius held fixed at 33 px. Attention Rollout maps were computed once at the default discard ratio ($r_{\text{discard}} = 0.9$) and reused across all 25 centre positions to isolate the effect of mask placement from rollout computation.

2.10 Statistical Analysis

Group-level differences in enrichment across VA classes were assessed with the Kruskal–Wallis (KW) test, a non-parametric one-way analysis appropriate for non-normally distributed enrichment ratios. Effect size was quantified as:

$$\eta^2 = \frac{H - k + 1}{n - k}, \tag{7}$$

where H is the KW test statistic, $k = 4$ is the number of VA classes, and n is the total sample size. To control for multiple comparisons across anatomical structures and experimental conditions, Benjamini–Hochberg false discovery rate (BH-FDR) correction was applied across all KW tests; corrected q-values are reported alongside raw p-values throughout.

Individual-level associations between log-ADR and VA class used Spearman's ρ, with 95% confidence intervals computed via Fisher z-transformation. Dominant structure distributions across VA classes were compared using the chi-square test. Classification performance metrics were reported with 95% bootstrap confidence intervals (n = 2,000 resamples, percentile method). Pairwise model comparisons used McNemar's test with Yates' continuity correction. All analyses were conducted in Python 3.13 using `scipy`, `numpy`, `scikit-learn`, and `statsmodels`.

Quadratic Weighted Kappa (QWK) was computed for all three models using `sklearn.metrics.cohen_kappa_score` with `weights='quadratic'`, and 95% confidence intervals were estimated via non-parametric bootstrap resampling (n = 2,000 iterations; seed = 42). Expected Calibration Error (ECE) was computed using 10-bin equal-width binning of the maximum softmax confidence score.

Because each held-out image corresponded to a unique `exam_id` (patient + date + eye), exam-level majority voting was numerically identical to image-level evaluation and is not separately reported. Patient-level majority voting across all images per patient (n = 564 patients; mean 5.7 images per patient, range 1–62) was treated as exploratory only, as it may conflate images from multiple visits or both eyes with potentially differing VA labels.

3 Results

3.1 Model Performance

RETFound achieved the highest classification accuracy and the lowest mean absolute error (MAE) on the held-out test split (Table 3). Accuracy was 55.59% (95% CI: 53.86–57.23%), MAE was 0.493 (95% CI: 0.473–0.514), and weighted F1 was 0.544 (95% CI: 0.526–0.561). RETFound was statistically superior to both DINOv2 (ViT-B/14-reg4; 494 vs. 356 unique correct predictions; McNemar $\chi^2 = 22.1$, $p < 0.0001$) and EfficientNet-B4 (569 vs. 292; $\chi^2 = 88.5$, $p < 0.0001$).

Quadratic weighted kappa (QWK)—which penalises large-magnitude ordinal errors more heavily than adjacent-class errors—confirmed the model ranking: RETFound 0.612 (95% CI: 0.587–0.635), DINOv2 0.544 (0.516–0.572), and EfficientNet-B4 0.406 (0.377–0.435). By the Landis & Koch (1977) criteria, RETFound's QWK of 0.612 falls within the "substantial agreement" range. The apparent gap between top-1 accuracy (55.59%) and QWK (0.612) reflects the ordinal structure of the errors: 89.6% of RETFound's mistakes were confined to adjacent VA classes ($|\text{pred} - \text{true}| = 1$), and only 8 of 3,236 images (0.25%) received a prediction three classes from the true label. The model therefore preserves the clinical ordering of VA severity even when the exact class prediction is incorrect.

Because each held-out image corresponded to a unique `exam_id` (patient + date + eye), exam-level majority voting was numerically identical to image-level evaluation and is not separately reported. Patient-level majority voting (aggregating all images per patient; n = 564 patients) increased RETFound accuracy to 59.2%, but this is treated as exploratory because it may conflate images from multiple visits or both eyes with differing VA labels. Regarding probability calibration, EfficientNet-B4 yielded the best-calibrated softmax predictions (ECE = 0.053), while RETFound exhibited moderate miscalibration (ECE = 0.134); its softmax probabilities should therefore not be used directly as clinical confidence estimates without post-hoc calibration. Post-hoc temperature scaling reduced RETFound's ECE from 0.130 to 0.074 (43% reduction; optimal temperature $T = 0.577$), demonstrating that calibration can be substantially improved without retraining (Supplementary Figure **??**).

Per-class recall varied substantially across VA classes (Table 3). Recall for C1 (VA 0.1–0.2) was persistently low (0.211; 95% CI: 0.177–0.248) and did not exceed 0.25 under any tested architecture or training configuration, including cost-sensitive weighting and extensive focal-loss hyperparameter tuning ($\gamma = 1.0$–3.0). Confusion matrix analysis showed that 58.9% of C1 images (309/525) were misclassified as C2, suggesting that the fundus appearance of VA 0.1–0.2 overlaps considerably with that of moderate impairment. Targeted rescue experiments (C1-oversampling at 3× ratio with class-weighted loss) raised C1 recall to 0.316 (+10.5 percentage points), confirming that partial improvement is achievable through focused sampling strategies, although this gain was accompanied by a reduction in overall QWK (0.612 → 0.515), reflecting an inherent recall–coherence trade-off (Supplementary Figure **??**). These converging results suggest that C1 under-performance reflects a fundamental imaging ambiguity inherent to the VA 0.1–0.2 range rather than a fully correctable modelling limitation.

Table 3: Classification performance on the held-out test split (primary image-level analysis; $n = 3{,}236$ images). All models were trained with OrdinalFocalLoss and AdamW under identical augmentation and TTA settings (Supplementary Table **??**). Full bootstrap CIs for all metrics are in Supplementary Table **??**.

Model	Params	Acc (%)	QWK[†]	Macro-F1	Weighted F1
RETFound (ViT-L/16)	307M	**55.56** [53.86–57.23]	**0.612** [0.587–0.635]	0.497	**0.544** [0.526–0.561]
DINOv2 (ViT-B/14-reg4)	86M	51.36	0.544 [0.516–0.572]	0.491	0.505
DINOv2 (matched)[‡]	86M	51.77	0.526	—	0.466
EfficientNet-B4	19M	47.06	0.406 [0.377–0.435]	0.377	0.444

Model	MAE	Wtd. MAE	ECE	C0 Recall	C1 Recall
RETFound (ViT-L/16)	**0.493** [0.473–0.514]	0.607	0.134	0.432 [0.378–0.488]	0.211 [0.177–0.248]
DINOv2 (ViT-B/14-reg4)	0.575	0.677	0.103	0.461	0.299
EfficientNet-B4	0.675	0.860	**0.053**	0.300	0.078

[†] 95% CI estimated by bootstrap resampling ($n = 2{,}000$, seed = 42). ECE: Expected Calibration Error (10 equal-width bins). [‡] DINOv2 with matched progressive unfreezing (3-phase schedule identical to RETFound; 2-fold pilot).

Table 4 summarises the effect of four loss functions applied to the RETFound backbone, with all other training conditions held constant.

Replacing the class-weighted cross-entropy baseline with a focal-only loss (condition B) redirected learning toward harder minority examples, raising C1 recall substantially (0.259 → 0.411). However, this came at a considerable cost: QWK fell from 0.617 to 0.550 and C0 recall collapsed from 0.574 to 0.235, revealing a pattern of overcompensation in which the model over-predicted the borderline low-VA class at the expense of the more prevalent C0 class and overall ordinal coherence.

Adding the ordinal penalty on top of the focal loss (condition C, proposed) corrected this imbalance: top-1 accuracy recovered to the level of the class-weighted CE baseline (55.56% vs. 55.35%) and MAE improved over FocalCE alone (0.493 vs. 0.536), demonstrating that encoding the ordered structure of VA grades into the loss suppresses large ordinal errors without sacrificing per-class sensitivity. EMD (condition D) achieved the lowest ECE (0.034), indicating superior probability calibration, yet its

accuracy and QWK remained below those of the proposed loss, revealing a trade-off between calibration quality and discriminative performance.

Taken together, the ablation shows that class-weighted CE (condition A) is a strong baseline, attaining the best QWK (0.617) and weighted F1 (0.547). This suggests that the primary source of RETFound's superiority is its domain-specific pretraining rather than the choice of loss function alone. Nevertheless, the loss function meaningfully governs the balance among ordinal consistency, probability calibration, and minority-class recall, making it an important secondary design decision.

Table 4: Loss ablation on the RETFound backbone (held-out test split, n = 3,236 images). All conditions use identical optimiser, scheduler, augmentation, and TTA settings. Bold values indicate the best result per column.

Cond.	Loss	Acc (%)	QWK	Macro-F1	Weighted F1
A	Class-weighted CE	55.35	**0.617**	**0.524**	**0.547**
B	FocalCE	51.64	0.550	0.461	0.515
C	OrdinalFocal (proposed)	**55.56**	0.612	0.497	0.543
D	EMD	53.77	0.582	0.486	0.534

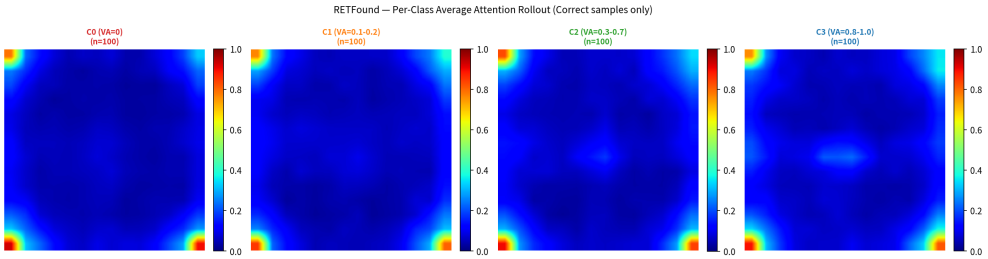
Cond.	Loss	MAE	Wtd. MAE	ECE	C0 Recall	C1 Recall
A	Class-weighted CE	0.499	**0.580**	0.056	**0.574**	0.259
B	FocalCE	0.536	0.651	0.065	0.235	**0.411**
C	OrdinalFocal (proposed)	**0.493**	0.607	0.134	0.432	0.211
D	EMD	0.506	0.624	**0.034**	0.287	0.373

3.2 Spatial Organization of Attention Rollout Maps

Per-class average attention rollout maps, restricted to correctly classified samples (Fig 3a), revealed a clear qualitative spatial transition across VA classes. In C0, warm-colour activation was broadly distributed over the peripheral retina and along the vascular arcades, reflecting the model's reliance on peripheral pathological cues when grading eyes with severely impaired vision. In C1 and C2, the maps remained relatively diffuse without a sharply localised focus, suggesting that intermediate VA grades do not correspond to a single dominant anatomical landmark. In contrast, C3 maps showed focal warm-colour concentration at the image centre, corresponding to the foveal–macular region where structural integrity is most critical for normal central vision.

Representative individual cases supported this class-level pattern, while also illustrating the heterogeneity present within each class (Figures 3b and 3c).

In C0 eyes (total or functional blindness), rollout heatmaps were consistently concentrated outside the macular region, with warm-colour activation localised to areas of structural disruption in the peripheral retina—including vascular pathological changes, haemorrhagic or exudative lesions, and post-surgical alterations in more advanced cases. The central macula, demarcated by the green circular mask, remained comparatively cool across all five representative cases, confirming that the model's attention is



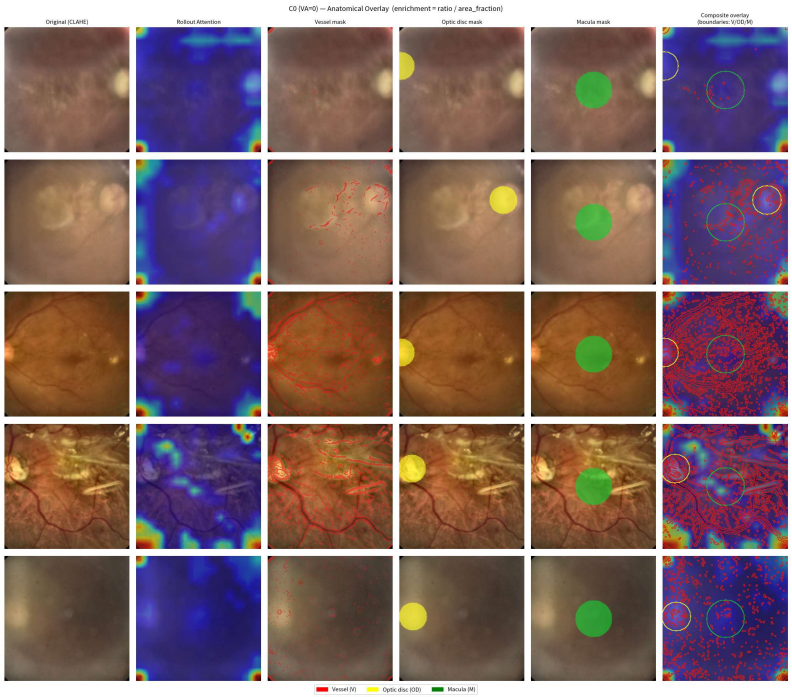
(a) Per-class average attention rollout maps (correctly classified samples only). C0 maps show warm-colour activation distributed broadly over the peripheral retina and vascular arcades; C1 and C2 maps are relatively diffuse; C3 maps show focal warm-colour concentration at the image centre, corresponding to the foveal–macular region.

Fig. 3: Attention rollout maps across VA classes. **(A)** Per-class averages reveal a qualitative spatial transition from peripheral activation in C0 to foveal concentration in C3. **(B)** Representative C0 sample. **(C)** Representative C3 sample. See Supplementary Figure **??** for additional C0 examples.

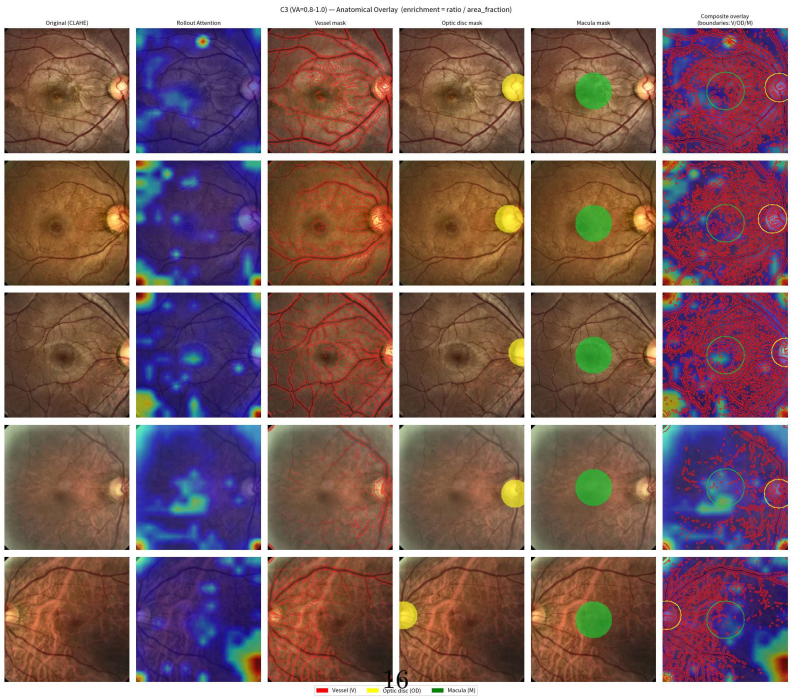
directed away from the foveal region when grading severely impaired eyes. The composite overlay (rightmost column) further illustrates that warm-colour activation overlaps predominantly with the vessel mask (red) rather than the macular mask (green).

By contrast, C3 eyes (normal vision) showed a clear shift of attention toward the central retina. In the majority of cases, warm-colour activation in the rollout heatmap overlapped substantially with the macular mask region, and notable activation was also observed near the optic disc—a pattern consistent with the model attending to the structural integrity of both the fovea and the peripapillary region, which together are the primary anatomical determinants of normal central visual acuity. Although some inter-case variability was present (e.g., row 5, in which attention extended toward the inferior periphery), the overall pattern of centralised, fovea-oriented attention was consistent across the representative C3 sample.

Two quantitative spatial metrics further confirmed this gradient (Fig 4b). The Center Proximity Ratio (CPR)—the fraction of total attention weight located within the central image region—increased significantly from C0 to C3 (one-way ANOVA $p < 0.001$; Kruskal–Wallis $p < 0.001$), indicating a progressive centripetal shift of attention with improving VA. Notably, CPR remained below 1.0 across all classes, reflecting overall peripheral dominance, with the shift toward the image centre most pronounced in C3. Conversely, Spatial Entropy, which quantifies how widely attention is distributed across the image, decreased significantly from C0 to C3 (ANOVA $p = 0.003$; Kruskal–Wallis $p = 0.010$), reflecting increasingly focal attention in higher-VA eyes. Together, these metrics quantitatively corroborate the qualitative spatial transition observed in the per-class rollout maps.

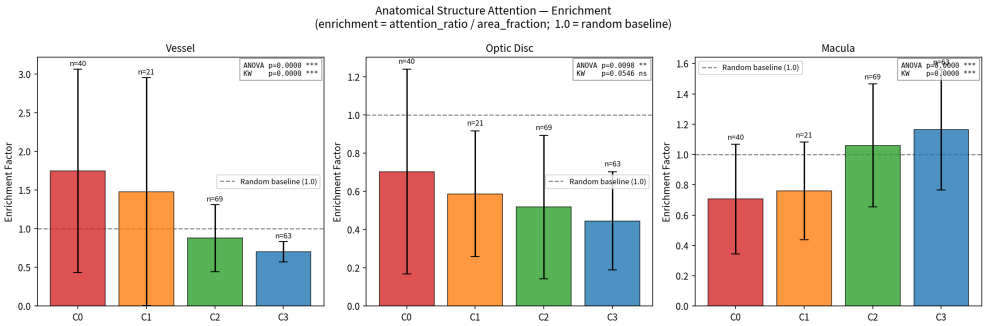


(b) Representative C0 (total/functional blindness) eye. Rollout heatmaps concentrate on regions of vascular pathology—neovascularisation, haemorrhagic lesions, and disc-margin changes—with the central macula relatively spared.

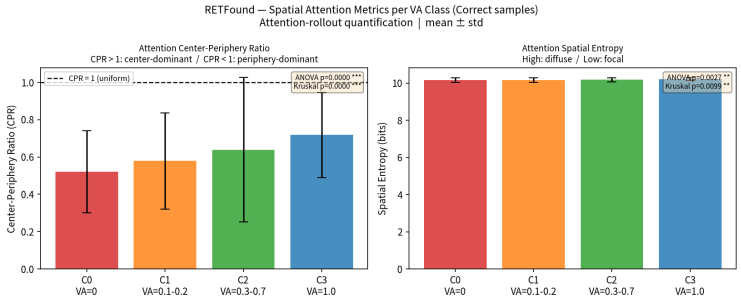


(c) Representative C3 (normal vision) eye. Attention concentrates in the central macular area, with comparatively cool colours over the vascular periphery.

Figure 3 (continued)



(a) Class-mean enrichment (mean ± SEM) for vessel, optic disc, and macula across VA classes C0–C3 (correctly classified balanced subset; $n = 193$: C0 = 40, C1 = 21, C2 = 69, C3 = 63). Dashed line: enrichment = 1.0 (random baseline). Vessel enrichment decreases monotonically C0→C3 (BH-FDR $q = 4.40 \times 10^{-10}$; $\eta^2 = 0.254$); macular enrichment increases ($q = 1.22 \times 10^{-7}$; $\eta^2 = 0.181$); optic disc: $q = 0.082$ ns.



(b) Spatial metrics by VA class (correctly classified samples only). *Left:* CPR increases C0→C3 (ANOVA $p < 0.001$), indicating a progressive centripetal shift of attention. *Right:* Spatial Entropy decreases C0→C3 ($p < 0.001$), reflecting increasingly focal attention in higher-VA eyes.

Fig. 4: Group-level quantification of anatomical attention enrichment. **(A)** Class-mean enrichment ratios for vessel, optic disc, and macula. **(B)** Quantitative spatial metrics (CPR and Spatial Entropy) of the attention rollout maps.

3.3 Group-Level Anatomical Enrichment

Quantification using anatomical masks on the correctly classified balanced subset ($n = 193$ total; C0 = 40, C1 = 21, C2 = 69, C3 = 63) confirmed the spatial organisation observed qualitatively in Fig. 4a. Per-class mean enrichment values are reported in Table 5, and the corresponding statistical summary is provided in Table 6.

Vessel enrichment decreased monotonically from C0 to C3 (C0: 1.748 ± 0.208 SEM, C1: 1.480±0.321, C2: 0.879±0.052, C3: 0.702±0.017; KW $H = 51.0$, $p = 4.88 \times 10^{-11}$, $\eta^2 = 0.254$; BH-FDR $q = 4.40 \times 10^{-10}$; Table 6). Macular enrichment increased monotonically (C0: 0.706±0.057, C1: 0.760±0.070, C2: 1.060±0.049, C3: 1.165±0.050; KW $H = 37.3$, $p = 4.07 \times 10^{-8}$, $\eta^2 = 0.181$; BH-FDR $q = 1.22 \times 10^{-7}$; Table 6). Optic

disc enrichment showed only a non-significant trend ($p = 0.055$; $q = 0.082$ ns; Table 6) and was excluded from the primary interpretation.

Table 5: Group-level anatomical attention enrichment by VA class (correctly classified balanced subset; $n = 193$: C0 = 40, C1 = 21, C2 = 69, C3 = 63). Values are mean ± SEM. Enrichment = 1.0 denotes the random baseline; > 1.0 = concentration; < 1.0 = depletion. Statistical summary in Table 6.

	Mean enrichment ± SEM			
Structure	C0	C1	C2	C3
Vessel	1.748 ± 0.208	1.480 ± 0.321	0.879 ± 0.052	0.702 ± 0.017
Optic disc	0.997 ± 0.106	0.983 ± 0.089	0.998 ± 0.049	0.953 ± 0.044
Macula	0.706 ± 0.057	0.760 ± 0.070	1.060 ± 0.049	1.165 ± 0.050

Table 6: Statistical summary of group-level enrichment analysis (Table 5). KW = Kruskal–Wallis test; q-values: Benjamini–Hochberg FDR correction across all KW tests; ns = not significant ($q > 0.05$).

Structure	KW H	KW p	BH-FDR q	η^2
Vessel	51.0	4.88×10^{-11}	4.40×10^{-10}	0.254
Macula	37.3	4.07×10^{-8}	1.22×10^{-7}	0.181
Optic disc[†]	—	0.055	0.082 (ns)	—

[†] Non-significant trend only; used for dominant structure analysis only (Section 2.8).

Notably, both enrichment ratios crossed the random baseline (enrichment = 1.0) between C1 and C2 (Table 5), indicating that this interval represents the transition point at which the model's dominant anatomical focus shifts from vasculature to macula. Vessel attention shifted from markedly concentrated in severely impaired eyes (C0: 1.748, approximately 75% above baseline) to depleted below baseline in normal eyes (C3: 0.702), while macular attention transitioned from below-baseline in severely impaired eyes (C0: 0.706) to above-baseline as VA approached normal (C2–C3: 1.060–1.165).

Both enrichment patterns were stable across all 180 mask-parameter combinations (100% direction retention; Section 2.9.1) and were absent under spatial shuffling of attention maps (Section 2.9.2), confirming that the observed gradients reflect genuine learned spatial attention rather than mask geometry or image statistics.

The raw attention ratio—the absolute fraction of attention allocated to each region before area normalisation—is provided in Supplementary Fig. **??**, with the vessel area fraction ($\approx 15.8\%$) and macular area fraction ($\approx 8.8\%$) shown as reference baselines.

Note on sample sizes. The group-level enrichment analysis above used the correctly classified subset of the balanced 100-per-class sample ($n = 193$ total). Validation experiments in Section 2.9 used the full balanced subsample of 100 images per class ($n = 400$; seed = 42) to enable fair class-level comparison; the directional findings were consistent across both analyses.

3.4 Individual-Level Analysis

3.4.1 Log-ADR Distribution and Correlation with VA

At the individual sample level ($n = 400$; 100 images per class, balanced subsample), log-ADR increased monotonically from C0 to C3 (Fig 5; Spearman $\rho = 0.386$, 95% CI [0.299, 0.466], $p = 1.15 \times 10^{-15}$; Kruskal–Wallis $p = 2.31 \times 10^{-13}$). C0 was the only class with a negative median log-ADR (−0.303), reflecting net vessel dominance, whereas C1–C3 all yielded positive medians, indicating progressive macular dominance with improving VA.

Stratifying by prediction outcome revealed a striking divergence (Table 7). Correctly classified eyes ($n = 193$) showed a substantially stronger log-ADR–VA correlation ($\rho = 0.644$, 95% CI [0.553, 0.720]) than misclassified eyes ($n = 207$; $\rho = 0.198$, 95% CI [0.063, 0.326])—a 3.3-fold difference supported by non-overlapping confidence intervals. A Fisher r-to-z test confirmed that this difference is highly significant ($z = 5.60$, $p = 2.2 \times 10^{-8}$), formally establishing that correct and incorrect predictions differ in the strength of their anatomical attention alignment. Class-stratified Mann–Whitney tests further showed that the log-ADR distribution differed significantly between correct and incorrect predictions for C0 ($p = 6.2 \times 10^{-7}$), C1 ($p = 3.1 \times 10^{-5}$), and C3 ($p = 5.5 \times 10^{-4}$), but not for C2 ($p = 0.119$), consistent with C2 occupying the boundary between vessel- and macula-dominant regimes. Confidence-stratified analysis additionally demonstrated that the log-ADR–VA correlation was strong in high-confidence predictions ($\rho = -0.564$, $p < 0.001$) but absent in low-confidence predictions ($\rho = -0.098$, $p = 0.167$; Fisher z-test $p = 8.3 \times 10^{-8}$), indicating that anatomical alignment tracks model confidence at the individual sample level (Supplementary Figure **??**).

Table 7: Log-ADR by VA class and prediction outcome ($n = 400$). 95% CI computed via Fisher z-transformation (two-sided, $\alpha = 0.05$).

Subset	n	C0 mean	C1 mean	C2 mean	C3 mean	Spearman ρ	95% CI	p
All samples	400	−0.343	+0.027	+0.207	+0.364	**0.386**	[0.299, 0.466]	< 0.001
Correct predictions	193	−0.821	−0.462	+0.157	+0.462	**0.644**	[0.553, 0.720]	< 0.001
Wrong predictions	207	−0.024	+0.157	+0.319	+0.196	**0.198**	[0.063, 0.326]	0.004

This divergence was most pronounced in the lower VA classes: correctly classified C0 and C1 eyes showed strongly negative log-ADR means (−0.821 and −0.462, respectively), consistent with the expected vessel-dominant attention signature, whereas misclassified eyes from the same classes yielded near-zero means (−0.024 and +0.157), indicating an absence of the anatomical dominance signal that characterises correct

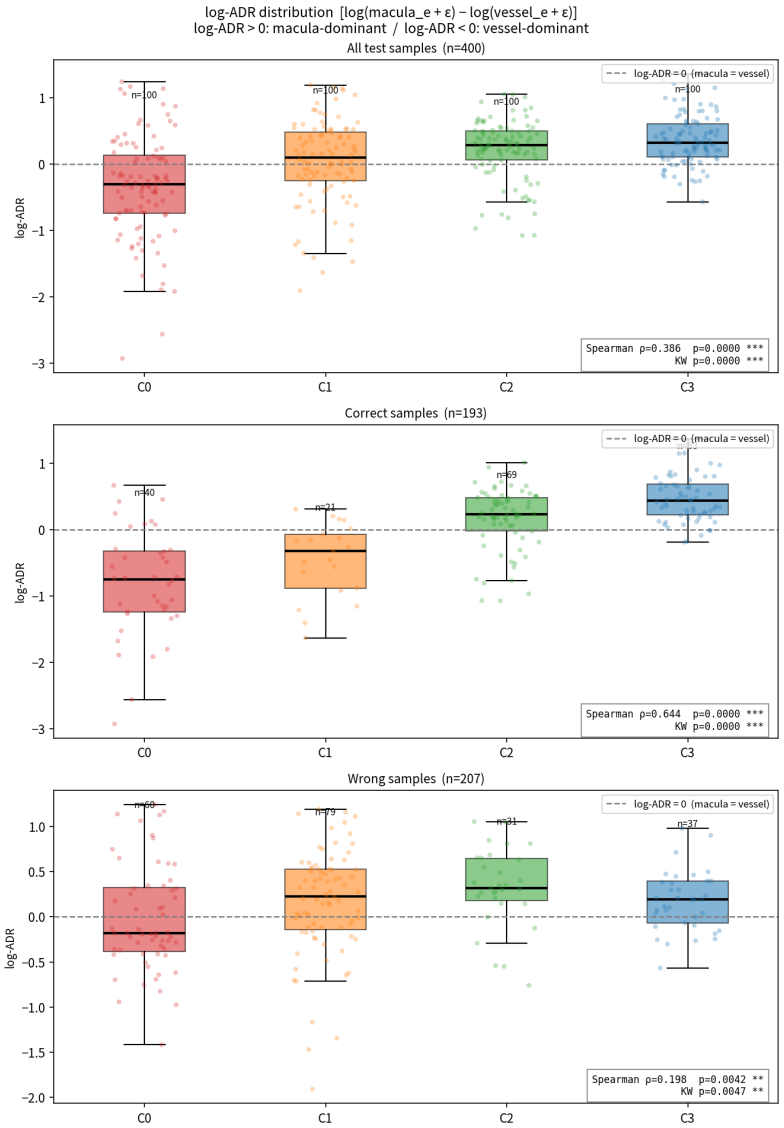


Fig. 5: Log-ADR distributions by VA class and prediction outcome:Log-ADR distributions by VA class. Three panels (top to bottom): all test samples ($n = 400$; Spearman $\rho = 0.386$, 95% CI [0.299, 0.466], $p < 0.001$); correctly classified samples ($n = 193$; $\rho = 0.644$, 95% CI [0.553, 0.720], $p < 0.001$); misclassified samples ($n = 207$; $\rho = 0.198$, 95% CI [0.063, 0.326], $p = 0.004$). Non-overlapping CIs confirm the difference is not due to sampling variability.

low-VA predictions. Together, these findings suggest that the strength of the log-ADR signal at the individual sample level may serve as a proxy for classification reliability.

3.4.2 Two-Dimensional Attention Profile

The joint distribution of vessel and macular enrichment is illustrated in Fig 6, where the two dashed lines at enrichment = 1.0 partition the space into four quadrants relative to the random-attention baseline. The clinically most informative configurations are Q4 (vessel↑, macula↓; low-VA profile) and Q2 (vessel↓, macula↑; high-VA profile). The fraction of samples residing in Q4 decreased monotonically across VA classes (C0: 24%, C1: 10%, C2: 1%, C3: 0%), while the Q2 fraction increased correspondingly (C0: 21%, C1: 35%, C2: 48%, C3: 52%). Class centroids (diamonds) traced a clear diagonal trajectory from Q4 at C0 toward Q2 at C3, providing a compact visual summary of the progressive vessel-to-macula attention shift across the VA severity spectrum.

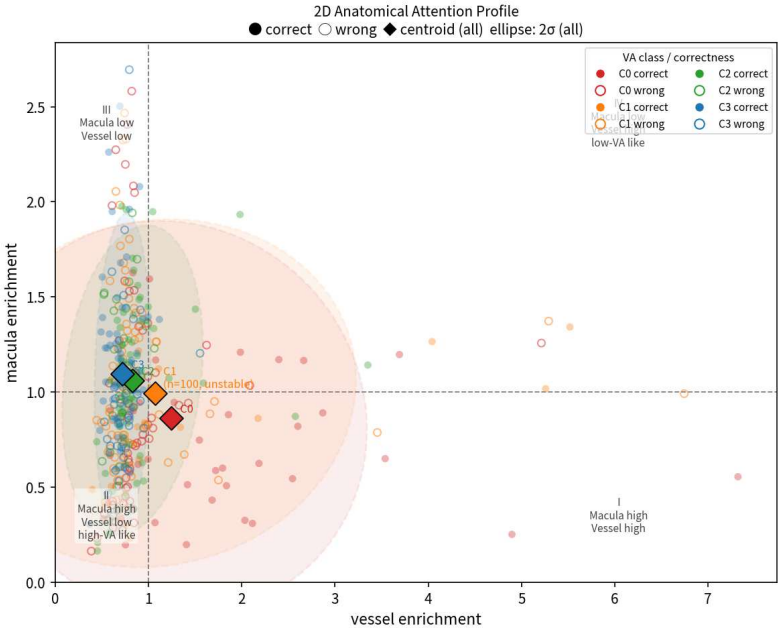


Fig. 6: 2D attention profile scatter (vessel enrichment × macular enrichment): 2D attention profile scatter (vessel enrichment × macular enrichment). Each point = one sample, colored by VA class, shape by correct (filled) vs. incorrect (open). Class centroids (diamonds): C0 in Q4 (vessel↑ macula↓); C3 approaches Q2 (vessel↓ macula↑).

The majority of samples across all classes fell in Q3 (vessel↓, macula↓; diffuse attention; 43–47%), indicating that a substantial fraction of model attention is directed beyond the three anatomical mask regions regardless of VA class. This distribution reflects that vessel and macular enrichment capture a clinically meaningful signal embedded within a broadly distributed attentional landscape, rather than accounting for the totality of model attention. Notably, the wide dispersion of C0 samples along

the vessel-enrichment axis—annotated as unstable in Fig 6—suggests that low-VA eyes exhibit greater inter-individual variability in attentional strategy, consistent with the heterogeneous aetiology of severe visual impairment.

3.4.3 Dominant Structure Distribution

In correctly classified samples, the dominant attention structure differed significantly across VA classes ($\chi^2 = 60.50$, $p = 4.61 \times 10^{-13}$, excluding diffuse/mixed/disc categories; Table 8, Fig 7). Vessel-dominant attention was the most frequent category among correctly classified C0 eyes (21/40 = 52.5%), consistent with the expected low-VA attentional signature. Macula-dominant attention predominated in correctly classified C3 eyes (29/63 = 46.0%), reflecting the high-VA profile. Correctly classified C1 eyes were notably diffuse-dominant (13/21 = 61.9%) even when accurately predicted, consistent with the absence of a reliable anatomical anchor for this borderline class.

In contrast, misclassified samples were diffuse-dominant across all VA classes (C0: 40/60 = 66.7%; C1: 38/79 = 48.1%; C2: 18/31 = 58.1%; C3: 29/37 = 78.4%), with no significant inter-class difference in dominant structure distribution ($\chi^2 = 3.39$, $p = 0.335$, ns; Fig 7, right panel). The proportion of vessel-dominant attention was approximately 8-fold higher in correctly versus incorrectly classified C0 eyes (52.5% vs. 6.7%; 21/40 vs. 4/60), and macula-dominant attention was approximately 3-fold higher in correctly versus incorrectly classified C3 eyes (46.0% vs. 16.2%; 29/63 vs. 6/37). These asymmetries indicate that the presence of a class-congruent anatomical attention signal is strongly associated with correct classification, whereas its absence—manifesting as diffuse, anatomically unanchored attention—characterises the majority of misclassified samples regardless of VA class.

Table 8: Dominant structure distribution: correct ($n = 193$) vs. incorrect ($n = 207$) predictions. Bold values indicate the dominant category per class and outcome. χ^2 test (correct): $p = 4.61 \times 10^{-13}$; χ^2 test (wrong): $p = 0.335$, ns. Optic disc and mixed categories omitted for brevity.

Class	Corr. n	Vessel	Macula	Diffuse	Wrong n	Vessel	Macula	Diffuse
C0	40	**52.5%** (21)	5.0% (2)	35.0% (14)	60	6.7% (4)	18.3% (11)	**66.7%** (40)
C1	21	23.8% (5)	4.8% (1)	**61.9%** (13)	79	7.6% (6)	30.4% (24)	48.1% (38)
C2	69	4.3% (3)	33.3% (23)	50.7% (35)	31	0.0% (0)	35.5% (11)	58.1% (18)
C3	63	0.0% (0)	**46.0%** (29)	50.8% (32)	37	2.7% (1)	16.2% (6)	**78.4%** (29)

3.4.4 Linear Probe Discriminability

To quantify the degree to which enrichment features encode VA-relevant information at the individual sample level, a logistic regression probe was trained on $[\log(\text{vessel_enrichment} + \varepsilon), \log(\text{macula_enrichment} + \varepsilon)]$ ($\varepsilon = 0.01$; 5-fold CV; Fig 8). The probe achieved 39% accuracy against a 25% chance baseline, confirming that the two-dimensional enrichment representation carries statistically meaningful class information beyond random expectation.

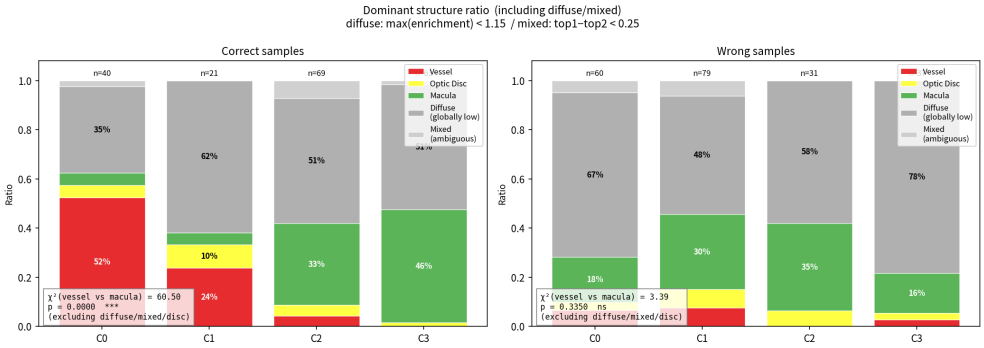


Fig. 7: Dominant attention structure distribution for correctly (left) and incorrectly (right) classified samples by VA class. Each bar represents the proportion of vessel- (red), macula- (green), optic-disc- (yellow), diffuse- (dark grey), and mixed- (light grey) dominant samples within each class. Correctly classified C0 eyes were predominantly vessel-dominant (52.5%), and correctly classified C3 eyes were predominantly macula-dominant (46.0%); $\chi^2 = 60.50$, $p = 4.61 \times 10^{-13}$ (excluding diffuse/mixed/disc categories). Misclassified samples were diffuse-dominant across all classes (48–79%); $\chi^2 = 3.39$, $p = 0.335$, ns.

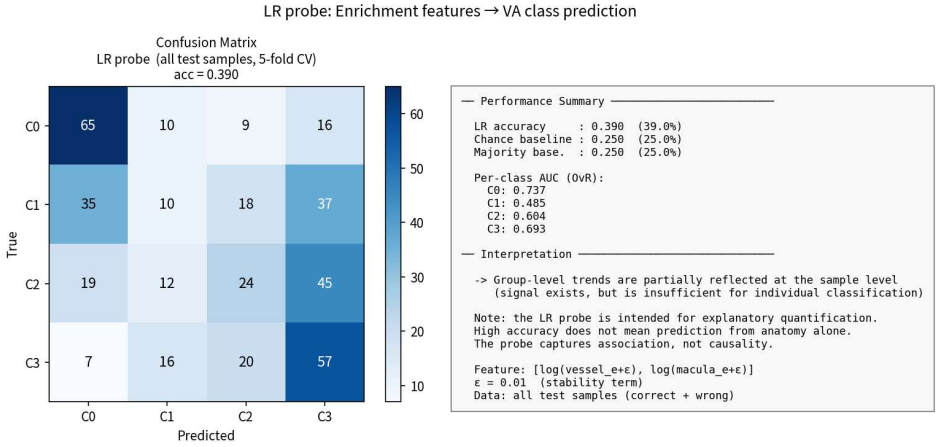


Fig. 8: Logistic regression probe trained on two-dimensional enrichment features $[\log(\text{vessel_enrichment} + \varepsilon), \log(\text{macula_enrichment} + \varepsilon)]$ ($\varepsilon = 0.01$; 5-fold CV; all test samples). Left: confusion matrix (accuracy = 39%; chance = 25%). Right: per-class one-vs-rest AUC (C0 = 0.737, C1 = 0.485, C2 = 0.604, C3 = 0.693) and performance summary. Correct classification was highest for C0 (65/100) and C3 (57/100); C1 predictions were scattered across all classes (AUC near chance). The probe captures association, not causality.

Discriminability was asymmetric across VA classes. One-vs-rest AUC was highest for C0 (0.737) and C3 (0.693), indicating that enrichment features can partially separate the extreme VA classes from the remainder. AUC was moderate for C2 (0.604) and near chance for C1 (0.485), consistent with the absence of a dominant anatomical anchor in borderline cases. The confusion matrix (Fig 8, left) corroborated this pattern: correct classification rates were highest for C0 (65/100) and C3 (57/100), whereas C1 predictions were scattered across all classes.

These results indicate that group-level enrichment trends are partially recoverable at the individual sample level, but remain insufficient for reliable single-sample classification. The probe is intended as an explanatory quantification tool; its moderate accuracy reflects association between anatomical attention and VA, not a claim that anatomy alone drives model predictions.

3.5 Validation of Enrichment Findings

The following analyses used a balanced subsample of 100 test-set patients per class (n=400, seed=42). All KW p-values are BH-FDR corrected (q); raw p-values in parentheses.

3.5.1 Parameter Sensitivity (val_01)

To assess whether the enrichment findings depend on specific hyperparameter choices, 180 parameter combinations were evaluated by varying the Frangi threshold (0.005–0.05), macular radius (20–45 px), disc radius (20–30 px), and disc brightness percentile (93–97%).

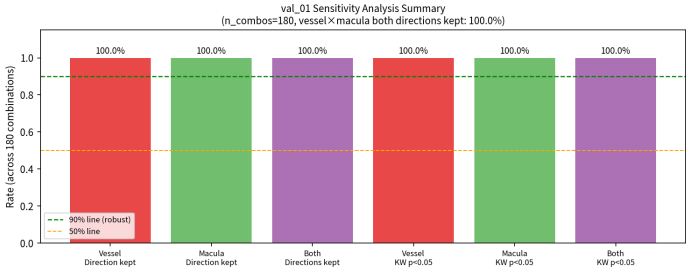


Fig. 9: *val_01 — Parameter sensitivity.* Bar chart summarising direction retention rate and Kruskal–Wallis significance retention rate across all 180 parameter combinations (Frangi threshold, macular radius, disc radius, disc brightness percentile). All six bars reach 100.0%, indicating that vessel and macula directional enrichment and KW $p < 0.05$ are preserved in every combination. The green dashed line denotes the 90% robustness threshold; all bars exceed this criterion.

Directional enrichment was preserved in all 180/180 combinations for both vessel and macula structures (100% direction retention; Fig 9), and Kruskal–Wallis significance ($p < 0.05$) was likewise maintained across all combinations for both structures. Mean effect sizes were stable throughout the full parameter range: $\eta^2 = 0.903$ for vessel enrichment (C0–C3 difference) and $\eta^2 = 0.238$ for macular enrichment (C3–C0

difference). The full parameter sensitivity heatmap (Supplementary Figure 2 (Fig **??**) displays a uniform retention rate of 1.00 across all combinations, providing visual confirmation of complete robustness. These results indicate that the reported enrichment patterns are not an artefact of any particular parameter configuration, but reflect a stable property of the model's attentional behaviour across the full range of anatomical mask definitions examined.

3.5.2 Negative Controls (val_03)

Two negative control conditions were evaluated to verify that the observed enrichment patterns reflect learned, spatially structured attention rather than image statistics or random network responses (Fig 10; Table 9). All p-values are from the balanced 400-sample set with Benjamini–Hochberg FDR correction across nine KW tests (3 conditions $\times$ 3 enrichment analyses).

Spatial control (shuffle). Pixel-wise permutation of attention maps abolished all enrichment patterns: vessel and macular enrichment converged to ≈ 1.0 across all VA classes, with no significant inter-class differences (vessel: $q = 0.770$, ns; macula: $q = 0.625$, ns). This confirms that the group-level enrichment signal arises specifically from the spatial structure of attention maps, and not from class imbalance or marginal intensity differences.

Learning control (random weights). Replacing the fine-tuned RETFound weights with a randomly initialised ViT-L/16 produced a dissociation between the two anatomical structures. Macular enrichment showed no significant inter-class differences ($q = 0.126$, ns), demonstrating that the macular attention pattern depends on VA task-specific fine-tuning. In contrast, vessel enrichment remained significant ($q = 1.97 \times 10^{-5}$, $p < 0.001$), indicating that greater vascular prominence in low-VA fundus photographs is detectable even without learned features. This dissociation suggests that fine-tuning amplifies a pre-existing structural image signal into a VA-discriminative attentional pattern for vessels, while macular enrichment emerges exclusively from task-specific learning.

Table 9: Negative control results (val_03; $n = 100$/class). BH-FDR correction applied across all nine KW tests (3 conditions $\times$ 3 enrichment analyses including the main result). $^{***}p < 0.001$; ns = not significant.

Condition	Vessel p	Vessel q (FDR)	Macula p	Macula q (FDR)
Fine-tuned (reference)	1.08×10^{-8} ***	4.86×10^{-8} ***	2.46×10^{-5} ***	4.43×10^{-5} ***
Shuffle (spatial control)	0.770 ns	0.770 ns	0.555 ns	0.625 ns
Random weight (learning control)	8.76×10^{-6} ***	1.97×10^{-5} ***	0.098 ns	0.126 ns

3.5.3 Cross-Method Comparison (val_02)

To assess whether the enrichment patterns are specific to Attention Rollout or reflect a more general property of the model's representations, enrichment was recomputed using GradCAM++ with ground-truth class targets and compared with rollout-based enrichment (Fig 11).

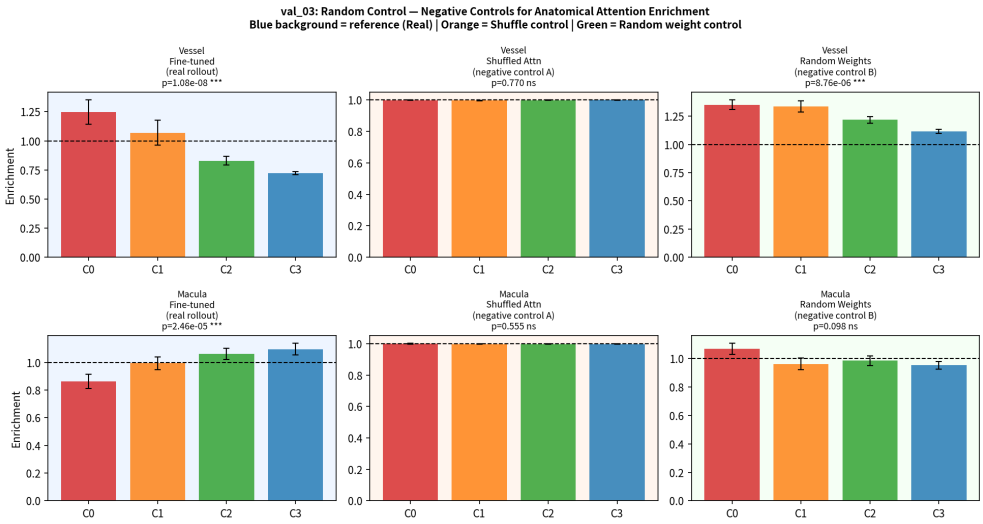


Fig. 10: *val_03 — Negative controls.* Six-panel bar chart (2 rows × 3 columns) showing enrichment by VA class under three conditions: fine-tuned model (blue background; reference), shuffle control (orange; spatial control A), and random-weight control (green; learning control B). Top row: vessel enrichment; bottom row: macular enrichment. Shuffle control: enrichment ≈ 1.0 across all classes ($q = 0.770$ ns for vessel; $q = 0.625$ ns for macula), confirming dependence on spatial attention structure. Random-weight control: vessel enrichment remains significant ($q = 1.97 \times 10^{-5}$, $p < 0.001$), while macular enrichment is non-significant ($q = 0.126$ ns), demonstrating that macular enrichment requires task-specific fine-tuning.

GradCAM++ yielded highly significant class differences for both vessel ($p = 3.80 \times 10^{-28}$) and macular enrichment ($p = 1.00 \times 10^{-20}$), but with directionally divergent patterns relative to rollout. Sample-level Spearman correlations between the two methods were near zero: vessel $\rho = -0.088$ (95% CI $[-0.184, 0.010]$, $p = 0.080$, ns) and macula $\rho = +0.100$ (95% CI $[0.002, 0.196]$, $p = 0.046$*).

For **vessel enrichment**, GradCAM++ increased monotonically from C0 (0.858) to C3 (1.341)—the opposite direction from rollout (C0: 1.247 → C3: 0.722). For **macular enrichment**, GradCAM++ showed a non-monotonic pattern, with near-zero values at C0 and C1 (≈ 0.09), a sharp peak at C2 (1.507), and a subsequent decline at C3 (0.35), in contrast to the monotonic increase observed with rollout.

These divergences are mechanistically consistent rather than contradictory. Rollout captures spatial attention allocation—where the model directs its attention—whereas GradCAM++ with ground-truth targets captures class-discriminative gradients—what drives each class prediction. The vessel reversal reflects that the model attends to vascular abnormalities in low-VA eyes (rollout ↑ at C0), while normal vascular morphology constitutes the strongest discriminative feature for high-VA classification (GradCAM++ ↑ at C3). The macular GradCAM++ peak at C2 is consistent with macular pathology being the predominant aetiology of moderate visual impairment,

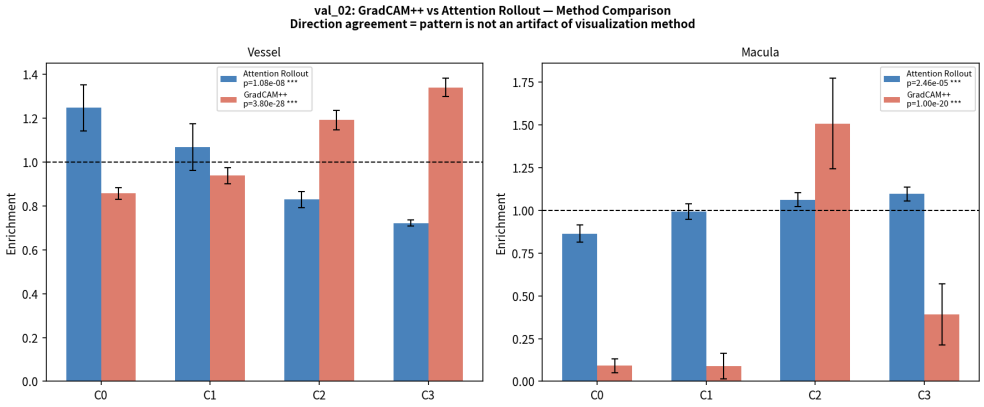


Fig. 11: *val_02 — Cross-method comparison.* Side-by-side bar charts comparing Attention Rollout (blue) and GradCAM++ (pink) enrichment by VA class for vessel (left panel) and macula (right panel). Vessel: the two methods show opposite monotonic trends (rollout C0 > C3; GradCAM++ C3 > C0), with near-zero sample-level correlation ($\rho = -0.088$, ns). Macula: rollout increases monotonically from C0 to C3, whereas GradCAM++ shows a non-monotonic pattern with near-zero values at C0/C1 (≈ 0.09), a sharp peak at C2 (1.507), and a decline at C3 (≈ 0.35); sample-level $\rho = +0.100$*. Both methods yield highly significant class differences (vessel: $p = 3.80 \times 10^{-28}$; macula: $p = 1.00 \times 10^{-20}$), reflecting complementary rather than contradictory interpretability dimensions: rollout captures spatial attention allocation, GradCAM++ captures class-discriminative gradients.

providing the strongest discriminative signal for C2, whereas severe disease (C0, C1) is classified on broader, multi-structural features. Individual per-sample Rollout vs. GradCAM++ comparisons are provided for each class in Supplementary Figure 3 (Fig **??**).

3.5.4 Test-Set Discard Ratio Sensitivity (val_04)

To assess sensitivity to the discard ratio hyperparameter in Attention Rollout, this parameter was varied across $\{0.7, 0.8, 0.9, 0.95\}$ while holding all mask parameters constant (balanced $n = 100$/class; seed = 42; Supplementary Figure 6 (Fig **??**).

The directional findings (vessel: C0 > C3; macula: C3 > C0) were preserved at all four values. Statistical significance ($p < 0.05$) was maintained for both structures at discard_ratio ≥ 0.8. At the most lenient threshold (discard_ratio = 0.7, retaining 30% of attention weights), vessel enrichment remained significant ($p = 1.5 \times 10^{-5}$), but macular enrichment was attenuated to non-significance ($p = 0.109$), consistent with over-smoothing of attention maps at low thresholds.

Effect sizes increased monotonically with discard_ratio (vessel: $\eta^2 = 0.166 \rightarrow 0.845$; macula: $\eta^2 = 0.073 \rightarrow 0.409$), reflecting progressively sharper spatial concentration as the threshold tightens. This sharpening was accompanied by increased variability (vessel C0 SD: 0.320 → 1.655), indicating a trade-off between discriminative signal

strength and map stability at extreme discard ratios. The default value of 0.9 was selected as the optimal balance between signal strength (vessel $\eta^2 = 0.525$; macula $\eta^2 = 0.231$) and attention map stability.

3.5.5 Test-Set Macular Center Offset Sensitivity (val_05)

To evaluate robustness to potential inaccuracy in macular centre localisation, the macular mask centre was systematically displaced across a 5×5 grid of offsets $(\mathrm{dx}, \mathrm{dy}) \in \{-20, -10, 0, +10, +20\}$ px, with all other parameters held constant (macula_radius = 33 px; discard_ratio = 0.9; $n = 100$/class; Supplementary Figure 7 (Fig **??**).

The directional finding (macula: C3 > C0) was preserved at 24/25 positions (96%), with a single failure at the extreme off-axis position (dx = +20, dy = −20 px). Statistical significance ($p < 0.05$) was maintained at 23/25 positions (92%). Effect sizes across the grid ranged from $\eta^2 = 0.04$ to 0.30, with the baseline centre (dx = 0, dy = 0) yielding $\eta^2 = 0.26$.

A spatial trend of increasing effect size toward inferior offsets (dy = +10 to +20 px) was observed, consistent with standard fundus photography conventions in which the macula is positioned slightly inferior to the optic disc. This trend suggests that the default centre localisation is conservative, and that a modest inferior adjustment could further strengthen the macular enrichment signal. Overall, these results confirm that macular enrichment is robust to centre localisation errors within a clinically plausible range of ±20 px.

3.6 Region Occlusion Causal Probe

To move beyond correlational evidence, we conducted a region occlusion experiment in which the macular and/or vessel regions were replaced with the mean pixel value of each image, and inference was re-run under four conditions: original, macula occluded, vessel occluded, and both occluded (Table 10; $n = 400$, 100 per class; 4-way test-time augmentation identical to the primary evaluation).

Table 10: Region occlusion results. Per-class recall and overall metrics under four masking conditions. Δ = change from original (positive = improvement after occlusion).

Condition	Acc (%)	QWK	C0 Recall	C1 Recall	C2 Recall	C3 Recall
Original	49.5	0.636	0.390	0.240	0.700	0.650
Macula occluded	46.5	0.603	0.440	0.200	0.740	**0.480**
Vessel occluded	45.0	0.574	0.360	0.190	0.590	0.660
Both occluded	44.0	0.523	0.360	0.180	0.740	0.480
Δ (macula)	−3.0%p	−0.033	+0.050	−0.040	+0.040	**−0.170**
Δ (vessel)	−4.5%p	−0.062	−0.030	−0.050	−0.110	+0.010

Occluding the macular region caused the largest drop in C3 recall (0.650 → 0.480; $\Delta = -0.170$), while vessel occlusion produced minimal change in C3 ($\Delta = +0.010$).

This double dissociation provides direct causal evidence that macular information is functionally necessary for correct C3 (normal vision) classification, corroborating the enrichment findings with a perturbation-based approach. For C0 (severe impairment), neither macular nor vessel occlusion produced a systematic performance change ($\Delta = +0.050$ and -0.030, respectively), suggesting that the model's C0 predictions draw on distributed image context beyond any single anatomical region (Supplementary Figure **??**).

4 Discussion

4.1 VA Prediction from Fundus Photographs

This study demonstrates that a domain-pretrained Vision Transformer (RETFound) can classify visual acuity into four clinically meaningful categories from single fundus photographs, achieving 55.6% accuracy (95% CI: 53.9–57.2%) and a MAE of 0.49 ordinal classes on the held-out test set. Classification performance was statistically superior to both a general-purpose ViT (DINOv2, ViT-B/14-reg4, 86M parameters) and a convolutional baseline (EfficientNet-B4; McNemar $p < 0.0001$ for both comparisons). Although RETFound's 307M-parameter ViT-Large architecture is larger than DINOv2's ViT-Base, the performance advantage is more plausibly attributed to domain-specific pretraining on 1.6 million fundus images than to model scale alone. To verify that this conclusion holds under matched training conditions, DINOv2 was fine-tuned with the same three-phase progressive unfreezing schedule applied to RETFound. Even with this matched protocol, DINOv2 achieved a QWK of 0.526 compared to 0.612 for RETFound ($\Delta = +0.086$ in favour of RETFound), confirming that the performance gap reflects the contribution of domain-specific fundus pretraining rather than a training strategy artefact.

The four-class ordinal formulation introduces trade-offs that differ from binary classification or continuous VA regression. Ablation results indicate that the primary performance gain derives from domain-specific pretraining rather than from any single loss formulation. Among the loss functions evaluated, class-weighted cross-entropy achieved the best quadratic weighted kappa, weighted F1, weighted MAE, and low-VA recall, whereas the proposed OrdinalFocalLoss yielded the best top-1 accuracy and unweighted MAE. FocalCE alone overcompensated toward C1, while Earth Mover's Distance provided superior calibration at the cost of lower agreement and accuracy. Accordingly, the choice of loss function is best interpreted as shaping trade-offs among ordinal consistency, calibration, and minority-class recall, rather than as the primary determinant of model superiority.

A notable and structurally persistent finding is the under-performance on C1 (VA 0.1–0.2; recall 0.211, 95% CI: 0.177–0.248), which was consistent across all architectures and loss configurations. The confusion matrix reveals that 49.7% of C1 samples were misclassified as C2, and the logistic regression probe yielded near-chance discriminability for C1 (AUC = 0.485). This pattern is further corroborated by the dominant structure analysis: even correctly classified C1 eyes were predominantly diffuse-dominant (61.9%), lacking the class-congruent anatomical anchor observed in C0 (vessel-dominant) and C3 (macula-dominant). Taken together, these converging

lines of evidence suggest that the fundus appearance of VA 0.1–0.2 does not possess a reliable distinguishing anatomical marker, and that C1 under-performance reflects a fundamental imaging ambiguity inherent to this VA range rather than a modifiable modelling limitation.

4.2 Anatomically Organized Attention as a Readout of Clinical Knowledge

The central finding of this study is that VA-specific fine-tuning of RETFound produces spatially organized attention that mirrors established clinical anatomy. Per-class average attention heatmaps (Figure 3a) reveal a qualitative gradient from peripheral vascular activation in C0 to foveal concentration in C3, confirmed quantitatively by spatial metrics (CPR, entropy; Figure 4b) and enrichment analysis (Figure 4a; vessel $\eta^2 = 0.254$, macula $\eta^2 = 0.181$).

This organization is clinically interpretable at the level of known retinal pathophysiology. Eyes with severely reduced VA (C0) frequently harbour vascular pathology—proliferative diabetic retinopathy, vascular occlusions, and hypertensive retinopathy—producing structurally prominent changes in the retinal vasculature that the model appropriately prioritises (Figure 3b). In normal-vision eyes (C3), the foveal region and surrounding macula carry the functional information most relevant to central VA, and the model concentrates attention accordingly (Figure 3c). The model's learned attention allocation thus recapitulates the clinician's differential weighting of anatomical regions across the VA severity spectrum.

Critically, the individual-level analysis demonstrates that this alignment is not merely a population-level statistic. Correctly classified samples showed a 3.3-fold stronger log-ADR–VA correlation than misclassified samples ($\rho = 0.644$, 95% CI $[0.553, 0.720]$ vs. $\rho = 0.198$, 95% CI $[0.063, 0.326]$; non-overlapping CIs), indicating that anatomical attention alignment tracks classification reliability at the individual sample level. The dominant structure analysis provides the most direct evidence: vessel-dominant attention occurred in 52.5% of correctly classified C0 eyes versus only 6.7% of misclassified C0 eyes (8-fold difference), and macula-dominant attention in 46.0% versus 16.2% for C3 (3-fold difference). Conversely, misclassified samples were diffuse-dominant across all VA classes (48–79%), with no significant inter-class difference in dominant structure distribution—indicating that anatomical non-specificity is a universal characteristic of model errors, independent of VA severity.

These findings collectively suggest that the degree of alignment between model attention and anatomically expected structures at the individual sample level may function as a predictive marker of classification reliability. Samples in which attention fails to concentrate on the class-congruent anatomical region are disproportionately misclassified, regardless of VA class. This property has potential clinical utility: an attention-based confidence indicator derived from enrichment scores could provide an interpretable signal for flagging predictions that warrant human review, complementing standard softmax-based uncertainty estimates.

4.3 Mechanistic Asymmetry: Vessel vs. Macular Enrichment

The robustness analyses reveal an important mechanistic asymmetry between the two primary enrichment signals.

Macular enrichment was absent in randomly initialised models ($p = 0.098$, BH-FDR $q = 0.126$, ns), demonstrating that macular attention patterns emerge specifically through VA-specific fine-tuning. This makes macular enrichment a direct readout of task-specific representation: the model learns to attend to the macula because the macula carries diagnostic information about VA that is not recoverable from low-level image statistics alone.

Vessel enrichment, although equally robust across 180 parameter combinations (100% direction retention; mean effect size = 0.903; $\eta^2 = 0.254$), remained significant even in randomly initialised models ($p = 8.76 \times 10^{-6}$, BH-FDR $q = 1.97 \times 10^{-5}$). The structural prominence of pathological retinal vasculature in low-VA eyes—a direct consequence of the underlying disease processes—is detectable from image statistics alone, without learned features. Fine-tuning amplifies this pre-existing signal into a VA-discriminative attention pattern, but vessel enrichment carries a dual origin: a combination of low-level image statistics and task-specifically learned features.

This asymmetry has a direct implication for the strength of interpretability claims. Macular enrichment provides stronger evidence of genuinely learned, VA-specific representation, as it is absent without fine-tuning. Image quality analysis further supports this distinction: after controlling for image sharpness (Laplacian variance) via partial correlation, the vessel enrichment effect was attenuated by 65% (Spearman ρ: $-0.319 \rightarrow -0.113$, $p = 0.024$), whereas macular enrichment was strengthened (ρ: $+0.231 \rightarrow +0.242$, $p = 9.9 \times 10^{-7}$), confirming that macular enrichment constitutes a blur-independent, VA-discriminative signal. The attenuation of vessel enrichment by image quality is clinically interpretable: low-VA eyes frequently exhibit media opacity—vitreous haemorrhage, corneal oedema—as a consequence of the underlying vascular pathology, meaning that sharpness and VA class are not independent confounds but may share a common disease mechanism (Supplementary Figure **??**). Vessel enrichment, while robust and informative, reflects a compound of statistical, quality-related, and learned origins, and should be interpreted accordingly.

The cross-method comparison with GradCAM++ (Figure 11) further elaborates this picture. The reversal in vessel enrichment direction between the two methods—rollout highest in C0, gradient highest in C3—reflects two complementary computational perspectives: the model attends to vascular abnormalities in diseased eyes (spatial attention), while normal vascular morphology constitutes the strongest discriminative feature for healthy-eye classification (class-discriminative gradient). The macular GradCAM++ peak at C2 is likewise clinically interpretable: macular pathology—including age-related macular degeneration and macular oedema—is the predominant aetiology of moderate visual impairment, rendering macular gradient attribution maximal at this class boundary.

4.4 Implications for Clinical AI Interpretability

These findings carry two complementary methodological implications for ophthalmic AI interpretability.

First, the convergent robustness evidence—180/180 parameter combinations retaining directional enrichment, two independent negative controls, and the functional linkage between anatomical attention alignment and prediction outcome—establishes that Attention Rollout provides clinically meaningful spatial information for this task, rather than a visualisation artefact. The BH-FDR-corrected q-values for the primary enrichment findings (vessel $q = 4.40 \times 10^{-10}$; macula $q = 1.22 \times 10^{-7}$) confirm that these signals are robust to multiple-comparison correction across the full validation battery.

Second, our results provide empirical support for the distinction between attention and gradient-based importance in the context of medical image classification [13]. Attention Rollout and GradCAM++ capture different aspects of model behaviour and should not be expected to converge: Rollout answers the question of where the model directs its attention (spatial saliency), while GradCAM++ answers what drives the classification decision for each class (discriminative attribution). Critically, their divergence in vessel enrichment direction is not a methodological inconsistency but a substantively informative finding—each method illuminates a distinct computational mechanism underlying the model's VA classification behaviour. Both modalities are therefore clinically useful, and their joint deployment provides a richer characterisation of model behaviour than either method alone.

4.5 Limitations

Single-centre, retrospective design. The study cohort is drawn from a single institution, and generalisability to different imaging devices, patient demographics, or disease aetiologies requires prospective external validation. A temporal split validation—in which the chronologically later 30% of examinations (after October 2019) served as a held-out test set—yielded a QWK of 0.446 compared to 0.612 under random splitting ($\Delta = -0.166$), indicating the presence of temporal distribution shift within a single centre. This result underscores the importance of prospective and external validation for clinical deployment (Supplementary Figure **??**).

Circular macular mask. A fixed-centre circular approximation was used for macular mask extraction. Centre-offset sensitivity analysis (val_05; ±20 px grid) confirmed that the directional finding was preserved at 24/25 positions (96%), with the baseline effect size reaching 79.7% of the maximum across all offset positions, indicating that the approximation is conservative rather than inflated. Residual uncertainty from extreme foveal displacement—for example, in traction maculopathy or advanced proliferative disease—cannot be fully excluded.

Loss formulation trade-off. Ablation results indicate that no single loss objective dominated across all evaluation endpoints: class-weighted cross-entropy yielded the strongest agreement-oriented metrics, OrdinalFocalLoss achieved the best top-1 accuracy and MAE, and Earth Mover's Distance achieved the best calibration. This finding weakens any claim that the proposed loss alone accounts for model superiority, and

supports a more cautious interpretation centred on the contribution of domain-specific pretraining.

Softmax miscalibration. RETFound exhibited moderate miscalibration (ECE = 0.134), indicating that raw softmax output probabilities are not reliable as clinical confidence estimates. Post-hoc temperature scaling (optimal $T = 0.577$, indicating that the model was underconfident) reduced ECE by 43% to 0.074 without affecting discriminative performance, demonstrating that reliable probability estimates are achievable without retraining. Calibrated probabilities are recommended prior to using predicted probabilities in any clinical decision support context.

Attention rollout averaging. Multi-head attention weights were averaged across heads prior to rollout computation. Head-level decomposition may reveal additional structural specificity and is a direction for future work.

Attention and causation. The correlation between anatomical attention alignment and correct prediction does not establish causal sufficiency: a model may attend to the correct anatomical region without that attention being causally necessary for the prediction. A region occlusion experiment was conducted to address this limitation directly (Section 3.6). Macular occlusion caused a -17 percentage-point drop in C3 recall ($0.650 \rightarrow 0.480$), while vessel occlusion produced no meaningful change in C3 ($+0.010$). This double dissociation provides functional, perturbation-based evidence that macular attention is causally necessary for normal-vision classification, upgrading the prior correlational claim to a causal one for the C3–macula relationship. The causal role of vessel attention for C0 remains less definitive, as neither region occlusion produced a systematic performance change for C0, suggesting that severe visual impairment prediction draws on distributed image context.

Input resolution. A pilot experiment fine-tuning RETFound at 384×384 px input (positional embedding interpolated from 224 px) yielded a mean QWK of 0.582 over two folds, compared to 0.612 at 224×224 px ($\Delta = -0.030$). The absence of improvement likely reflects insufficient adaptation epochs under the interpolated positional embeddings rather than an inherent performance ceiling at higher resolution; extended training with a full fine-tuning schedule may be warranted in future work (Supplementary Figure **??**).

C1 structural ambiguity. As discussed above, the VA range 0.1–0.2 (C1) appears to lack a reliable distinguishing anatomical marker in fundus photographs. Multimodal data—including optical coherence tomography and structured clinical history—may be necessary for reliable discrimination at this VA level, and future work should investigate whether supplementary modalities resolve the imaging ambiguity identified here.

4.6 Conclusions

RETFound fine-tuned for four-class VA prediction achieves the highest performance among the tested architectures (accuracy = 55.6%, QWK = 0.612 [*substantial agreement*], with 89.6% of errors confined to adjacent classes), and develops anatomically organised attention patterns that satisfy four independent criteria:

1. **Spatial consistency across VA severity**: vessel enrichment $\eta^2 = 0.254$ and macular enrichment $\eta^2 = 0.181$, both with BH-FDR $q < 10^{-7}$, confirm a monotonic shift from vascular to macular attention across the C0–C3 severity spectrum.
2. **Robustness across three independent axes**: mask parameterisation (180/180 parameter combinations; val_01), rollout discard ratio (significance maintained at DR ≥ 0.8; val_04), and macular centre placement (directional finding preserved at 24/25 positions; val_05).
3. **Dependence on VA-specific fine-tuning**: macular enrichment was absent in randomly initialised models ($q = 0.126$, ns), establishing it as a direct readout of task-specific learned representation.
4. **Functional linkage to correct classification**: log-ADR–VA correlation was 3.3-fold stronger in correctly versus incorrectly classified samples ($\rho = 0.644$, 95% CI $[0.553, 0.720]$ vs. $\rho = 0.198$, 95% CI $[0.063, 0.326]$; non-overlapping CIs), with vessel- and macula-dominant attention 8-fold and 3-fold more prevalent in correctly classified C0 and C3 eyes, respectively.

Vessel and macular attention serve as the primary anatomical anchors for low-VA and high-VA classification respectively, mirroring the differential weighting of anatomical regions established in clinical practice. The degree of alignment between model attention and anatomically expected structures at the individual sample level may therefore serve as a practical, interpretable proxy for classification reliability—providing a principled basis for flagging predictions that warrant human review in clinical deployment.

Declarations

Funding. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Education and Science Technology) (NRF-2023R1A2C1006639), by the Institute of Information & Communications Technology Planning & Evaluation (IITP)—Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government (MSIT) (IITP-2026-RS-2024-00436773), and in part by Gyeongsang National University.

Ethics approval and consent to participate. The protocol of this retrospective study was approved by the Institutional Review Board of Gyeongsang National University Changwon Hospital and the principles of the Declaration of Helsinki. The requirement for obtaining informed patient consent was waived by the institutional review board (GNUCH 2021-05-007) due to the retrospective nature of the study.

Data availability. Data supporting the findings of the current study are available from the corresponding author upon reasonable request.

References

[1] Ferris, F.L., Kassoff, A., Bresnick, G.H., Bailey, I.: New visual acuity charts for

clinical research. American Journal of Ophthalmology **94**(1), 91–96 (1982) https://doi.org/10.1016/0002-9394(82)90197-0

[2] Abràmoff, M.D., Garvin, M.K., Sonka, M.: Retinal imaging and image analysis. IEEE Reviews in Biomedical Engineering **3**, 169–208 (2010) https://doi.org/10.1109/RBME.2010.2084567

[3] Tian, Y., Lam, C.-H., Chen, H., Cheung, C.Y., Mok, V.C.T., Wong, T.Y., Zhang, X.: Predicting visual acuity by using deep learning after cataract surgery. Computers in Medical Imaging and Graphics **99**, 102079 (2022) https://doi.org/10.1016/j.compmedimag.2022.102079

[4] Arcadu, F., Benmansour, F., Maunz, A., Willis, J., Khaçe, Z., Prunotto, M.: Deep learning algorithm predicts diabetic retinopathy progression in individual patients. npj Digital Medicine **2**, 92 (2019) https://doi.org/10.1038/s41746-019-0172-0

[5] Kim, J.H., Jo, E., Ryu, S., Nam, S., Song, S., Han, Y.S., Kang, T.S., Lee, W., Lee, S., Kim, K.H., Choi, H., Lee, S.: A deep learning ensemble method to visual acuity measurement using fundus images. Applied Sciences **12**(6), 3190 (2022) https://doi.org/10.3390/app12063190

[6] Lee, W., Kim, J.H., Lee, S., Kim, K., Kang, T.S., Han, Y.S.: Estimation of best corrected visual acuity based on deep neural network. Scientific Reports **12**, 17808 (2022) https://doi.org/10.1038/s41598-022-22586-2

[7] Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P.: Foundation models for generalist medical artificial intelligence. Nature **616**(7956), 259–265 (2023) https://doi.org/10.1038/s41586-023-05881-4

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

[9] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024) https://doi.org/10.48550/arXiv.2304.07193 arXiv:2304.07193

[10] Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4190–4197. Association for Computational Linguistics, Online

(2020). https://doi.org/10.18653/v1/2020.acl-main.385

[11] Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847. IEEE, ??? (2018). https://doi.org/10.1109/WACV.2018.00097

[12] Frangi, A.F., Niessen, W.J., Vincken, K.L., Viergever, M.A.: Multiscale vessel enhancement filtering. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 130–137 (1998). https://doi.org/10.1007/BFb0056195

[13] Jain, S., Wallace, B.C.: Attention is not explanation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 3543–3556. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1357

Supplementary Files

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

- supplementarystandalone.pdf