

Supplementary Material

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

Kim et al.

Supplementary Tables

Supplementary Table S1. Training conditions for all evaluated models

Parameter	RETFound (ViT-L/16)	DINOv2 (ViT-B/14-reg4)	EfficientNet-B4
Pretrain data	1.6M fundus (MAE)	LVD-142M (DINO)	ImageNet-1K
Parameters	307M	86M	19M
Loss	OrdinalFocalLoss ($\gamma = 2$, $\lambda = 1$)	OrdinalFocalLoss ($\gamma = 2$, $\lambda = 1$)	OrdinalFocalLoss ($\gamma = 2$, $\lambda = 1$)
Optimizer	AdamW (wd = 1×10^{-4})	AdamW (wd = 1×10^{-4})	AdamW (wd = 1×10^{-4})
Scheduler	ReduceLROnPlateau	ReduceLROnPlateau	ReduceLROnPlateau
Augmentation	Full (Section 2.2)	Full (Section 2.2)	Full (Section 2.2)
TTA	4-crop mean	4-crop mean	4-crop mean
Backbone unfreezing	Progressive (3 phases)	Full fine-tune	Full fine-tune
Head LR	1×10^{-3}	1×10^{-3}	1×10^{-3}
Backbone LR	$2 \times 10^{-5} \rightarrow 1 \times 10^{-5}$	2×10^{-5}	2×10^{-5}
Early stopping	Val weighted-F1 (patience = 20)	Val weighted-F1 (patience = 20)	Val weighted-F1 (patience = 20)

All three models were trained under identical conditions with respect to loss function, data augmentation (Section 2.2 of main text), test-time augmentation (4-crop mean), optimiser, learning rate scheduler, and early stopping criterion (validation weighted-F1, patience = 20), on the same held-out test split (564 patients, 3,236 images). Progressive unfreezing across three phases was applied to RETFound only, to mitigate catastrophic forgetting of domain-specific representations acquired during pretraining on 1.6 million fundus images. The backbone learning rate for RETFound was additionally decayed from 2×10^{-5} to 1×10^{-5} across training phases, whereas DINOv2 and EfficientNet-B4 used a fixed backbone learning rate of 2×10^{-5} .

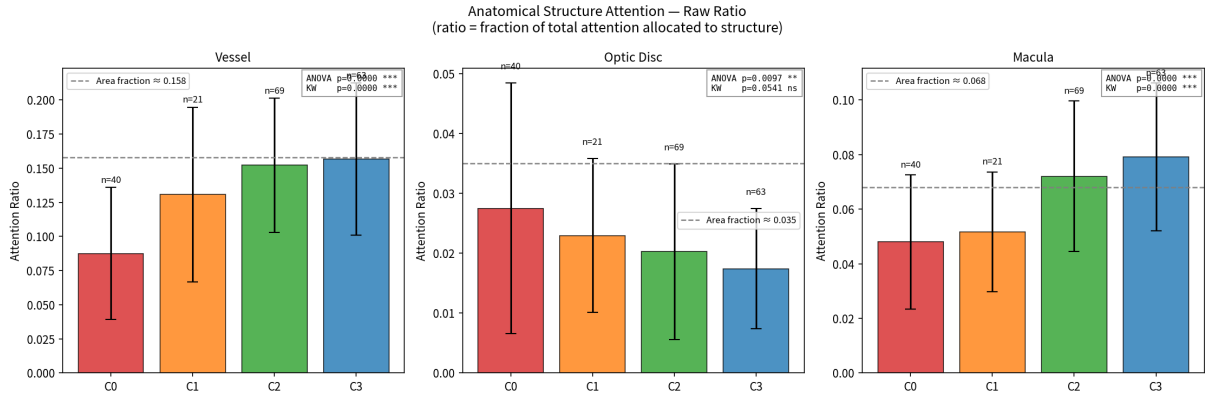
Supplementary Table S2. Full bootstrap confidence intervals for all evaluation metrics

Metric	RETFound	DINOv2	EfficientNet-B4
Accuracy (%)	55.56 [53.86–57.23]	51.36	47.06
QWK	0.612 [0.587–0.635]	0.544 [0.516–0.572]	0.406 [0.377–0.435]
Macro-F1	0.497	0.491	0.377
Weighted F1	0.544 [0.526–0.561]	0.505	0.444
MAE ↓	0.493 [0.473–0.514]	0.575	0.675
Weighted MAE ↓	0.607	0.677	0.860
ECE ↓	0.134	0.103	0.053
C0 Recall	0.432 [0.378–0.488]	0.461	0.300
C1 Recall	0.211 [0.177–0.248]	0.299	0.078

95% CIs estimated via bootstrap resampling ($n = 2,000$; seed = 42; percentile method). Bold values indicate the best-performing model per metric. For metrics where lower values are preferable (MAE, Weighted MAE, ECE), the ↓ symbol is noted in the metric name. Note that EfficientNet-B4 achieved the lowest ECE (0.053), indicating superior calibration relative to RETFound (ECE = 0.134), despite lower classification performance across all other metrics. Full bootstrap CIs for DINOv2 and EfficientNet-B4 are omitted for brevity and are available upon request.

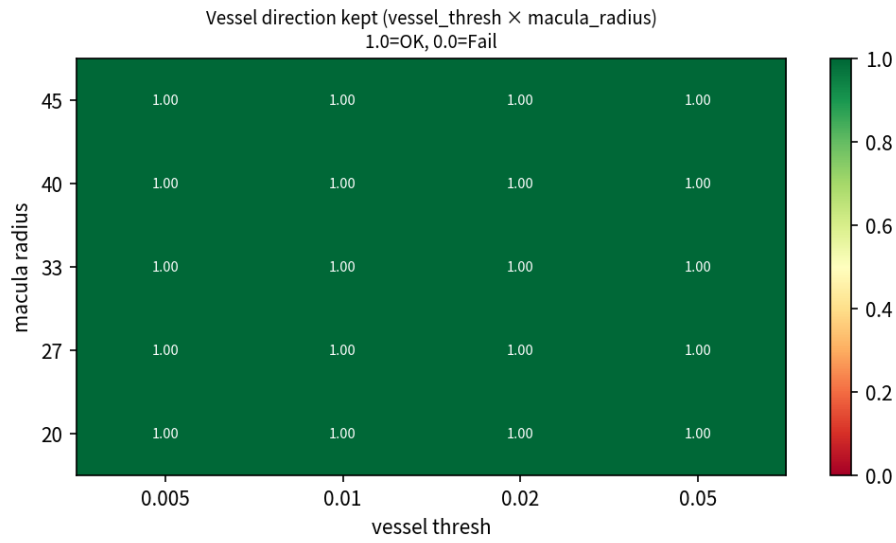
Supplementary Figures

Supplementary Figure S1. Raw attention ratio



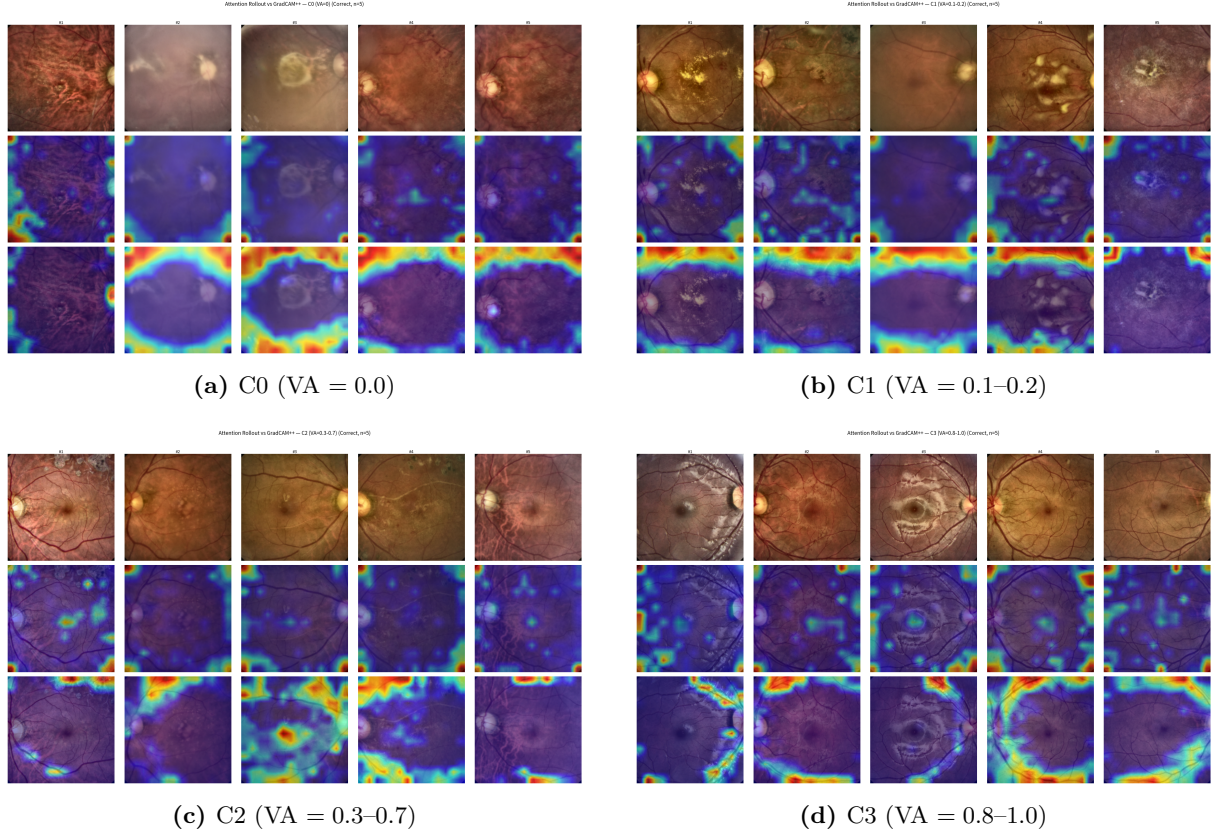
1: Supplementary Figure S1. Raw attention ratio (numerator only, before area normalization) for vessel, optic disc, and macula across VA classes C0–C3. Same layout as main Figure 5 but showing absolute attention fractions. Dashed lines indicate area fraction baselines (vessel $\approx 15.8\%$, macula $\approx 8.8\%$).

Supplementary Figure S2. Parameter sensitivity heatmap



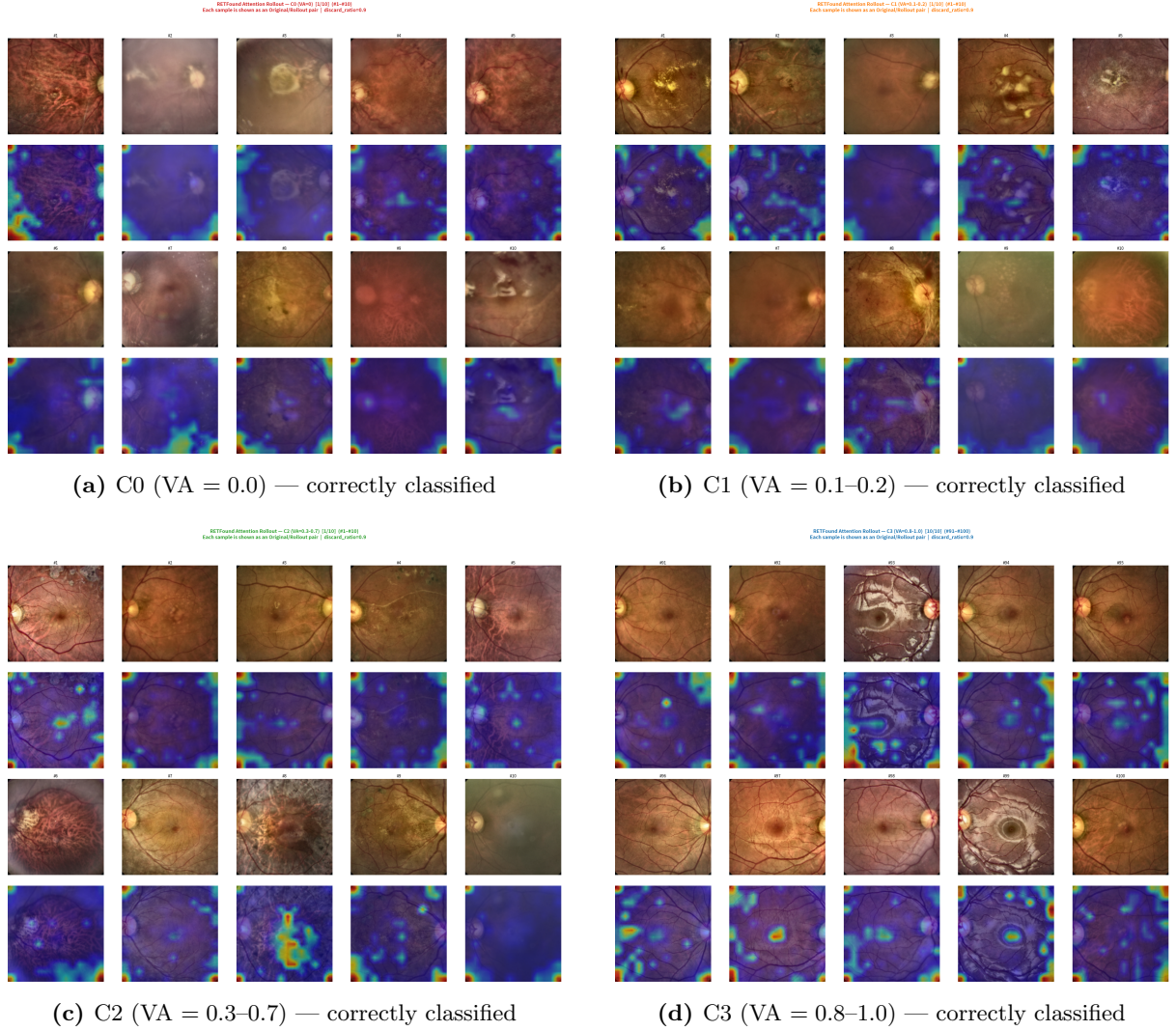
2: Supplementary Figure S2. Full parameter sensitivity heatmap showing vessel direction retention rate = 1.00 across all 180 parameter combinations (Frangi threshold $\times$ macular radius grid). The uniform green field confirms complete robustness. The macula direction heatmap is identical.

Supplementary Figure S3. Individual-level Rollout vs. GradCAM++ comparisons



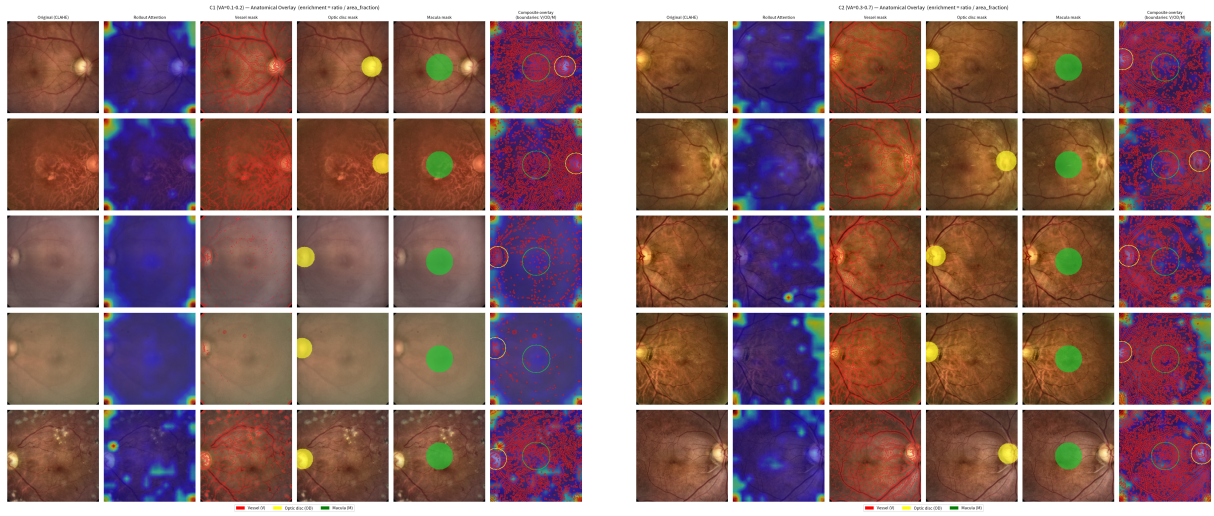
3: Supplementary Figure S3. Individual-level Rollout vs. GradCAM++ comparisons. Five representative correctly classified examples per VA class. Each panel shows three rows: CLAHE-preprocessed fundus image; Attention Rollout heatmap overlay; GradCAM++ heatmap overlay. The directional divergence (rollout vessel enrichment highest in C0; GradCAM++ vessel enrichment highest in C3) is visible in individual samples.

Supplementary Figure S4. Individual attention rollout examples per VA class



4: Supplementary Figure S4. Representative correctly classified examples with attention rollout overlays, anatomical mask overlays (vessel in red, macula in green, optic disc in yellow), and composite visualization per VA class. Ten correctly classified examples were generated for each class; representative subsets are shown.

Supplementary Figure S4b. Anatomical overlay examples — C1 and C2

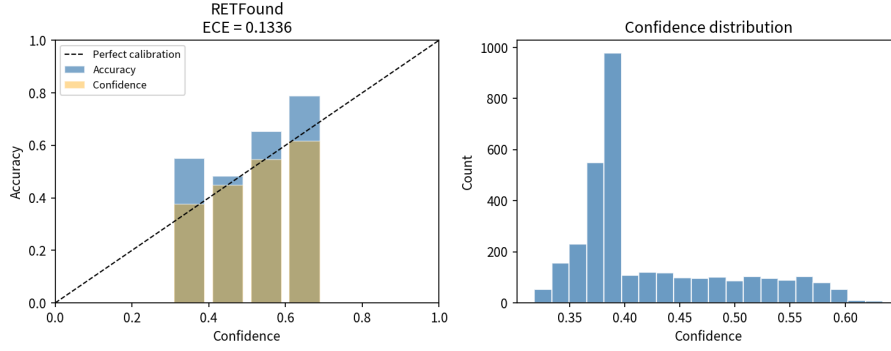


(a) C1 (VA = 0.1–0.2) — anatomical mask overlays

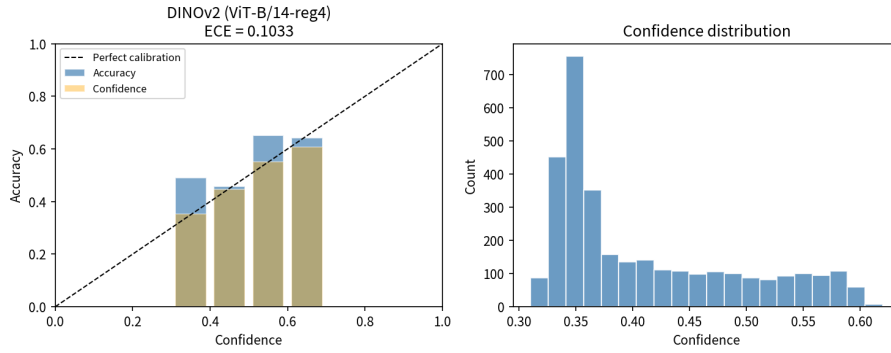
(b) C2 (VA = 0.3–0.7) — anatomical mask overlays

5: Supplementary Figure S4b. Representative anatomical overlay examples for intermediate VA classes C1 and C2. Same format as main Figures 3 and 4: CLAHE-preprocessed fundus image; Roll-out heatmap overlay; Vessel mask (red); Macula mask (green circle); Optic disc mask (yellow circle); Composite overlay. C1 and C2 maps show relatively diffuse attention patterns compared to the strongly vessel-dominant C0 and macula-dominant C3.

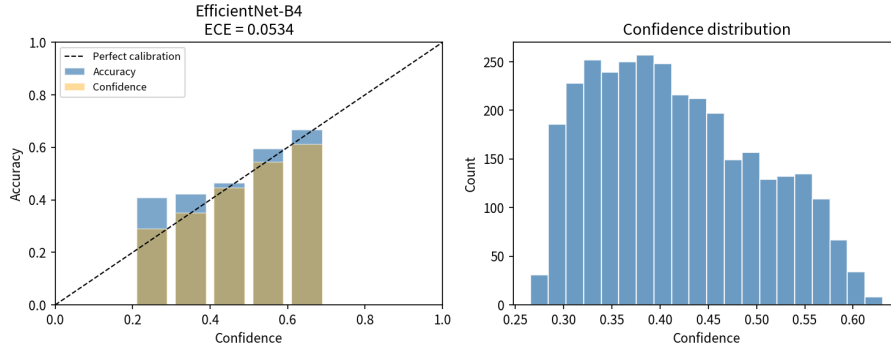
Supplementary Figure S5. Reliability diagrams



(a) RETFound (ECE = 0.134)



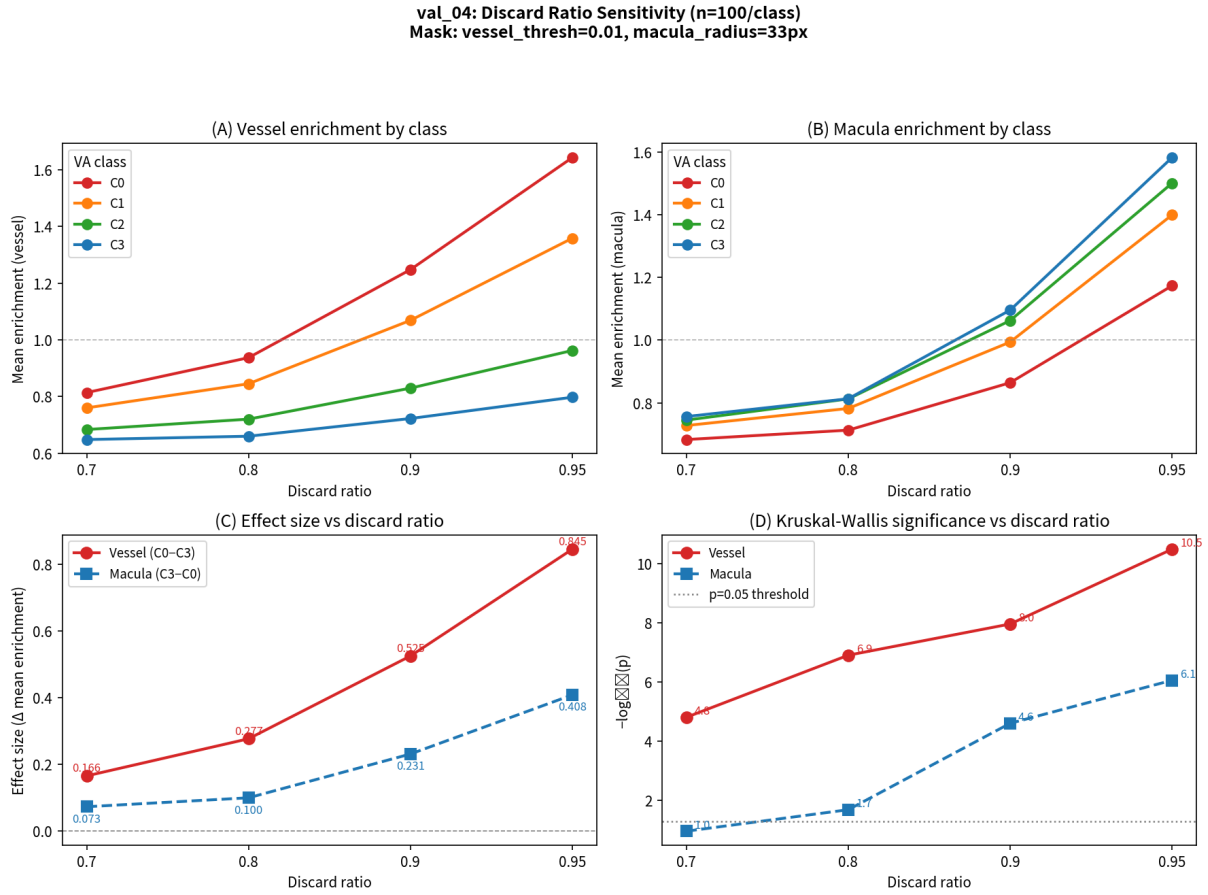
(b) DINOv2 (ECE = 0.103)



(c) EfficientNet-B4 (ECE = 0.053)

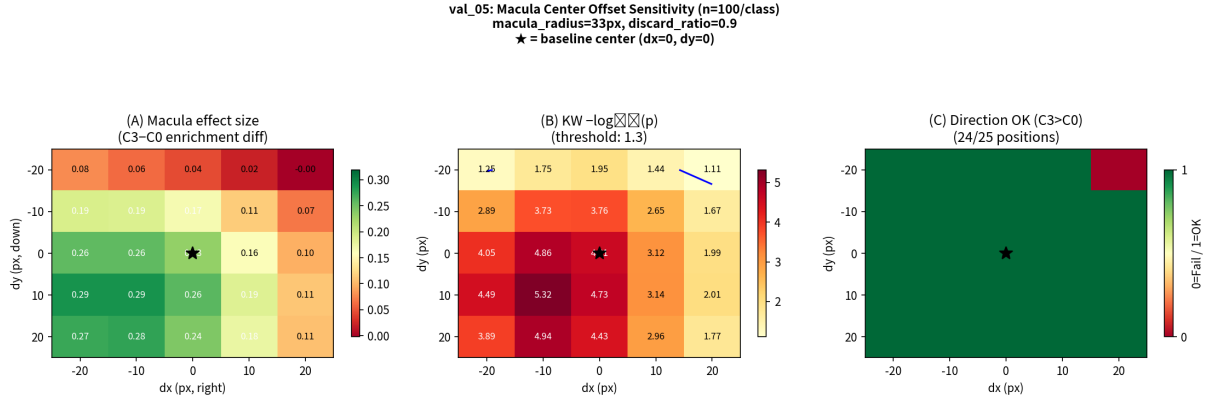
6: Supplementary Figure S5. Reliability diagrams on the held-out test split. Confidence versus empirical accuracy (10-bin equal-width ECE) for the three primary architecture comparisons. EfficientNet-B4 is best calibrated (ECE = 0.053); RETFound shows moderate miscalibration (ECE = 0.134) and should not be used directly as a clinical confidence estimate without temperature scaling.

Supplementary Figure S6. Discard-ratio sensitivity analysis



7: Supplementary Figure S6. Test-set discard-ratio robustness analysis. 2×2 panel showing, across discard ratios $\{0.7, 0.8, 0.9, 0.95\}$: (A) vessel enrichment by VA class; (B) macular enrichment by VA class; (C) effect size (η^2) for vessel (C0–C3 difference, red) and macula (C3–C0 difference, blue); (D) Kruskal–Wallis significance ($-\log_{10}(p)$) for vessel (red) and macula (blue), with the $p = 0.05$ threshold shown as a dotted line. Directional findings are preserved at all values; statistical significance is maintained at discard ratio ≥ 0.8 for both structures. The default value of 0.9 provides the optimal balance between signal strength and map stability.

Supplementary Figure S7. Macular center-offset sensitivity analysis

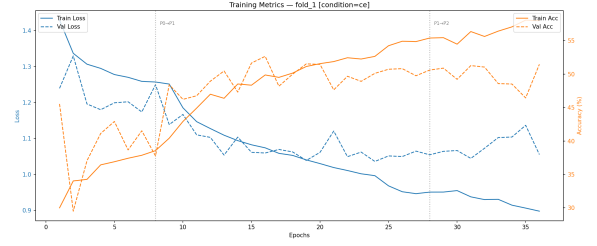


8: Supplementary Figure S7. Test-set macular center-offset robustness analysis. Three-panel heatmap across a 5×5 offset grid $(dx, dy) \in \{-20, -10, 0, +10, +20\}$ px; the baseline centre ($dx = 0, dy = 0$) is marked with a star (★). (A) Effect size (η^2 , C3-C0 enrichment difference): values range from 0.04 to 0.30; effect size increases toward inferior offsets ($dy = +10$ to $+20$ px). (B) Kruskal-Wallis significance ($-\log_{10}(p)$; threshold = 1.3): significance maintained at 23/25 positions (92%). (C) Direction retention ($C3 > C0$; green = retained, red = failed): directional finding preserved at 24/25 positions (96%); single failure at $dx = +20, dy = -20$ px. The spatial trend toward inferior offsets is consistent with standard fundus photography conventions.

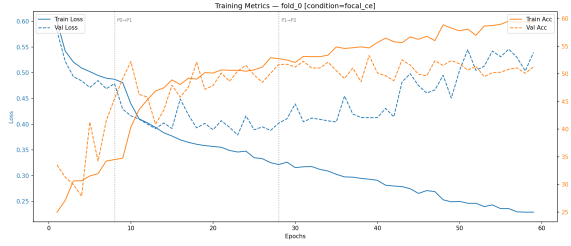
Supplementary Figure S8. Loss ablation training curves



(a) Class-weighted CE — Fold 0



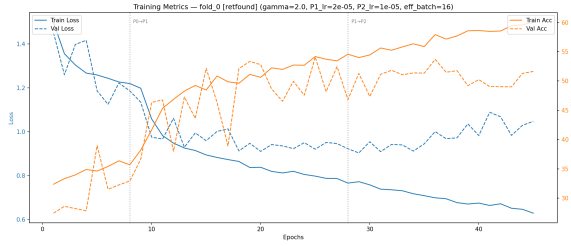
(b) Class-weighted CE — Fold 1



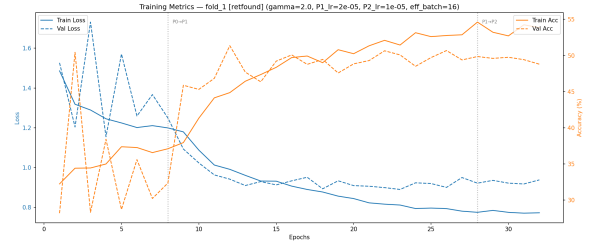
(c) FocalCE — Fold 0



(d) FocalCE — Fold 1



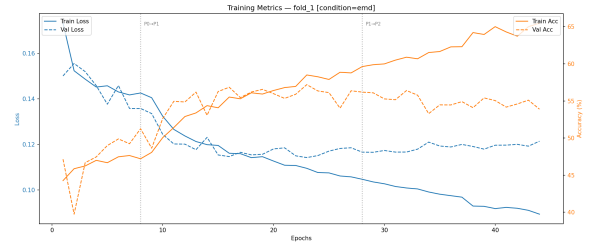
(e) OrdinalFocalLoss — Fold 0



(f) OrdinalFocalLoss — Fold 1



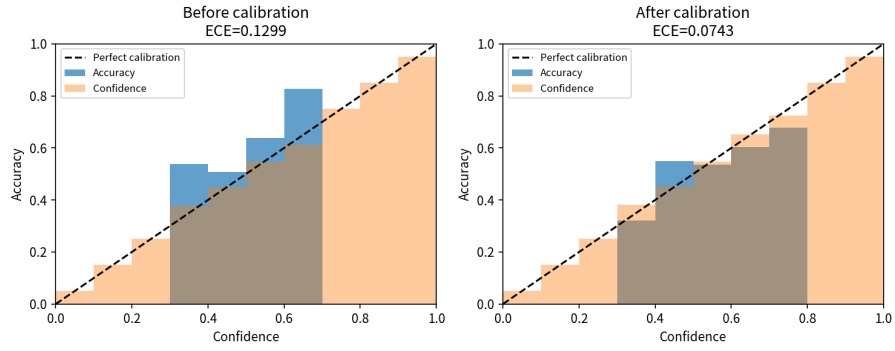
(g) EMD — Fold 0



(h) EMD — Fold 1

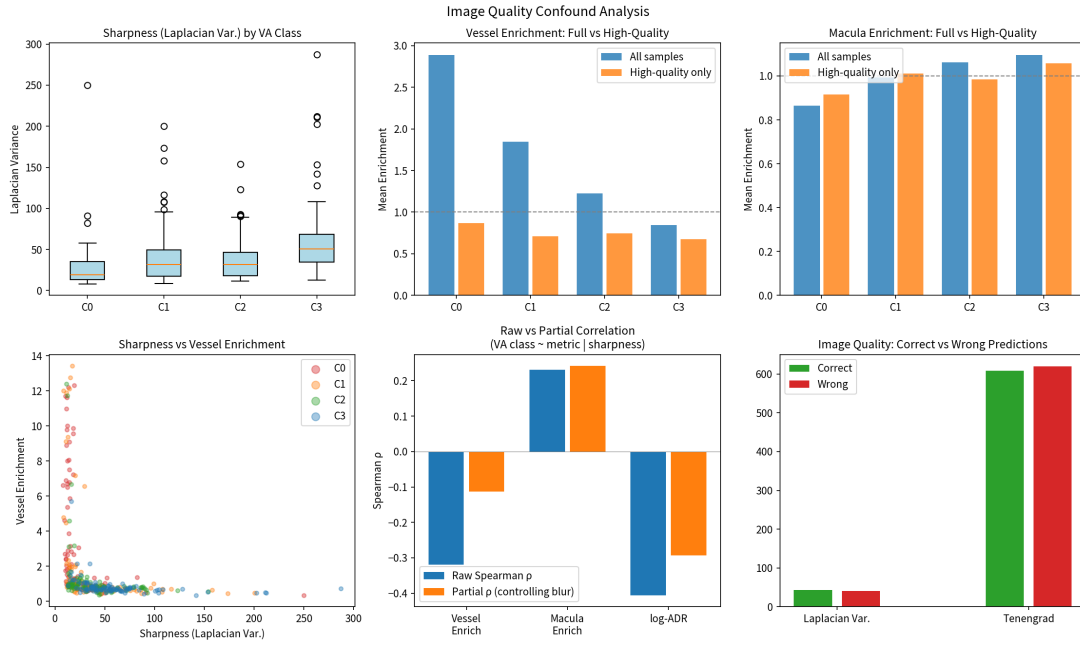
9: Supplementary Figure S8. Loss ablation training curves on the RETFound backbone. Fold-wise training and validation metrics for all four loss conditions (Class-weighted CE, FocalCE, OrdinalFocalLoss, EMD) across both cross-validation folds. OrdinalFocalLoss and class-weighted CE show similar convergence behavior, whereas FocalCE exhibits greater validation metric variability.

Supplementary Figure S9. Temperature scaling calibration



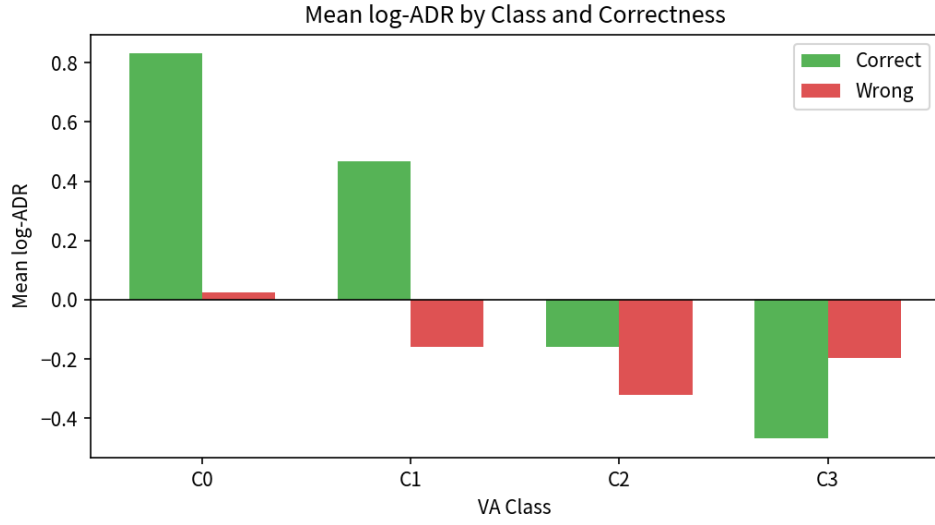
10: Supplementary Figure S9. Reliability diagram for RETFound before and after temperature scaling ($T = 0.577$). Expected Calibration Error (ECE) was reduced from 0.130 to 0.074 (43% reduction) without affecting discriminative performance. $T < 1.0$ indicates that the model was underconfident prior to calibration.

Supplementary Figure S10. Image quality confound analysis

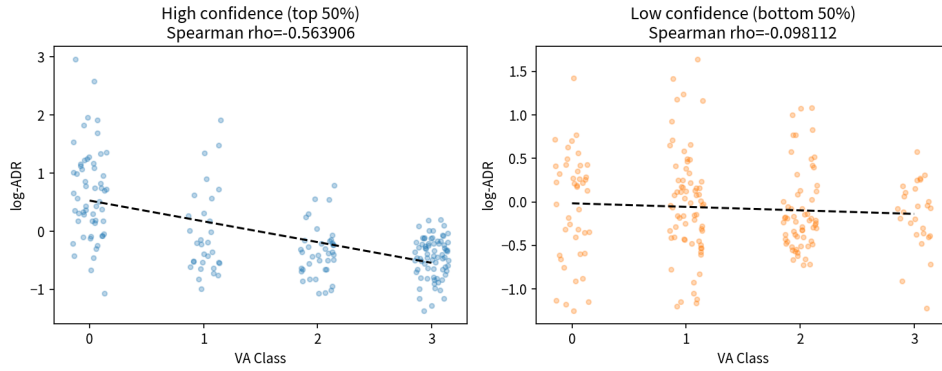


11: Supplementary Figure S10. Image quality confound analysis. Sharpness (Laplacian variance) was significantly lower in C0 images (Kruskal–Wallis $H = 70.3$, $p = 3.8 \times 10^{-15}$). After controlling for sharpness via partial correlation, vessel enrichment was attenuated by 65% ($\rho : -0.319 \rightarrow -0.113$, $p = 0.024$), whereas macular enrichment was strengthened ($\rho : +0.231 \rightarrow +0.242$, $p = 9.9 \times 10^{-7}$), confirming that macular enrichment is image-quality independent.

Supplementary Figure S11. Fisher r -to- z : log-ADR correlation analysis



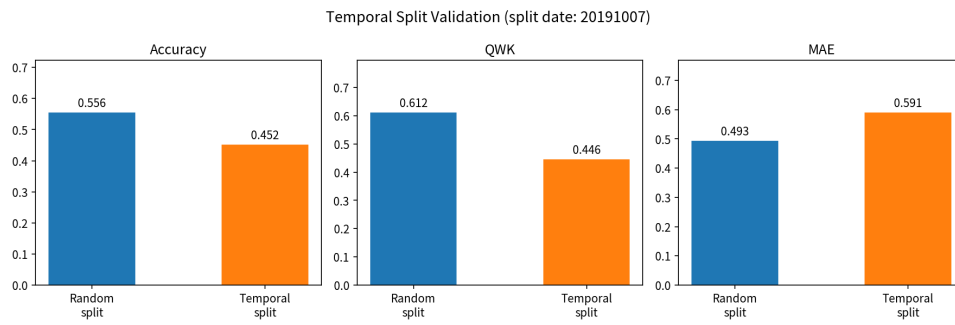
(a) Class-stratified log-ADR correlation (correct vs. incorrect predictions per class)



(b) Confidence-stratified log-ADR-VA correlation (high vs. low confidence)

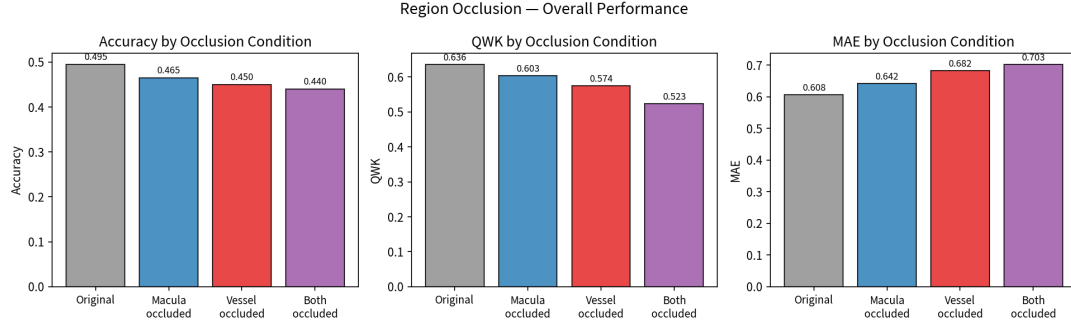
12: Supplementary Figure S11. Formal statistical analysis of log-ADR correlations. (a) Class-stratified Mann-Whitney tests confirm significant correct/incorrect log-ADR differences for C0, C1, and C3 ($p < 0.001$), but not C2. (b) High-confidence predictions show strong log-ADR-VA correlation ($\rho = -0.564$), whereas low-confidence predictions show no significant correlation ($\rho = -0.098$; Fisher z -test $p = 8.3 \times 10^{-8}$).

Supplementary Figure S12. Temporal split validation

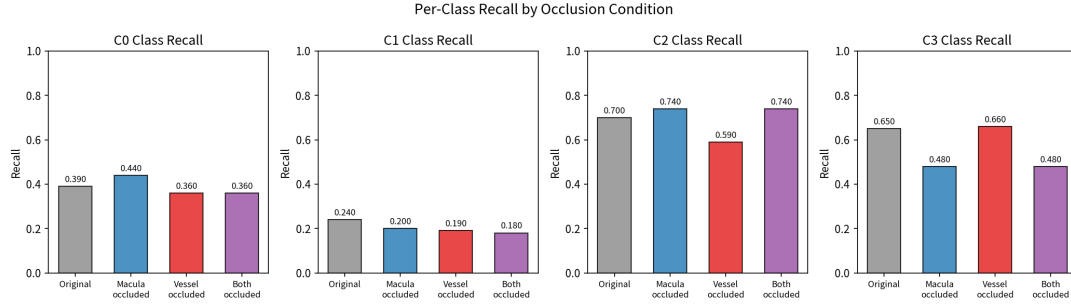


13: Supplementary Figure S12. Temporal split validation. Random split (standard evaluation) vs. chronological split (training on examinations before October 2019; testing on later examinations). QWK dropped from 0.612 to 0.446 ($\Delta = -0.166$) under the temporal split, indicating temporal distribution shift within the single-centre cohort.

Supplementary Figure S13. Region occlusion causal probe



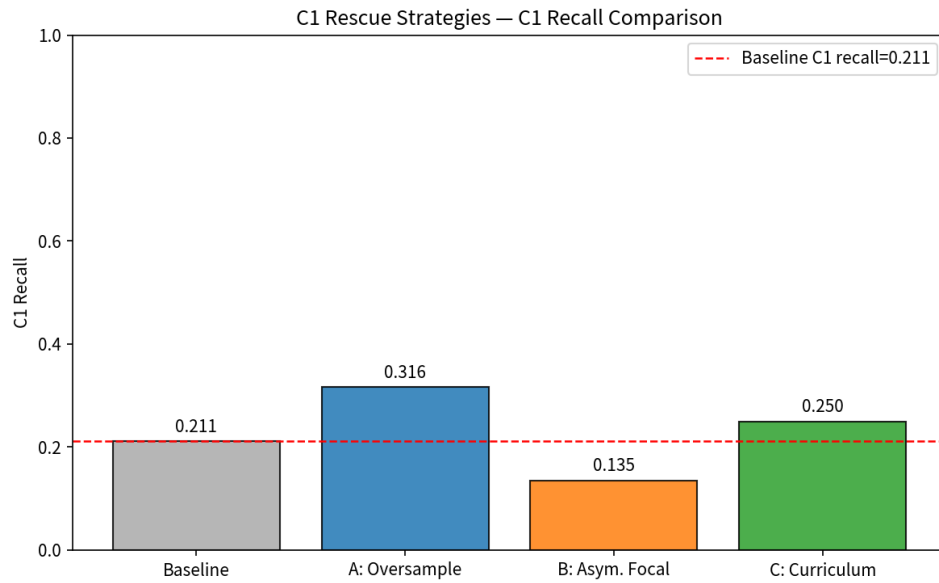
(a) Overall accuracy, QWK, and MAE under four occlusion conditions



(b) Per-class recall under each occlusion condition

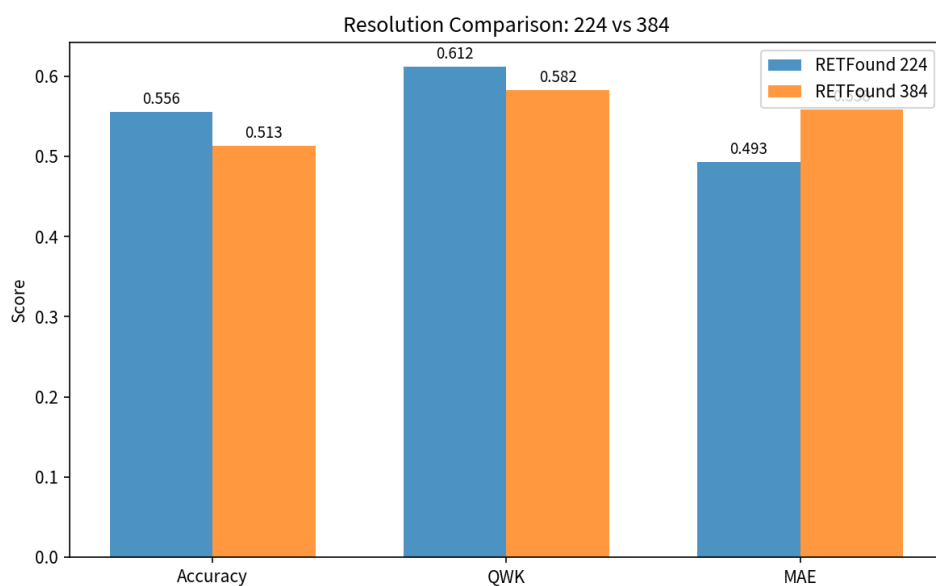
14: Supplementary Figure S13. Region occlusion causal probe ($n = 400$; 4-way TTA). Macular occlusion caused the largest drop in C3 recall ($0.650 \rightarrow 0.480$; $\Delta = -0.170$), providing direct causal evidence that macular information is functionally necessary for normal-vision classification. Vessel occlusion produced minimal change in C3 ($\Delta = +0.010$).

Supplementary Figure S14. C1 rescue strategy comparison



15: Supplementary Figure S14. C1 recall under three rescue strategies. Strategy A ($3\times$ oversampling with class-weighted loss) achieved the highest C1 recall (0.316; baseline 0.211), at the cost of reduced overall QWK (0.515 vs. 0.612). Strategy C (curriculum: binary pre-training then 4-class fine-tuning) provided a more balanced improvement (C1 recall 0.250; QWK 0.566). Strategy B (asymmetric focal loss) reduced C1 recall below baseline.

Supplementary Figure S15. High-resolution pilot (384 px)



16: Supplementary Figure S15. High-resolution pilot results. RETFound fine-tuned at 384×384 px (positional embedding interpolated from 224 px) achieved a mean QWK of 0.582 over two folds, compared to 0.612 at 224×224 px ($\Delta = -0.030$). Performance did not improve at higher resolution under the current training configuration, likely due to insufficient positional embedding adaptation epochs.