RETINAROBUST: BENCHMARKING ARTIFACT ROBUSTNESS AND ADAPTATION IN RETINAL FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retinal foundation models support diverse diagnostic tasks through the analysis of routine, non-invasive images of the eye. Nevertheless, these images are subject to a range of imaging artifacts and variations across clinical settings. We benchmark six retinal foundation models on 17 diagnostic tasks to understand their robustness to artifacts and color removal. All models exhibit performance degradation with the loss of image quality and chromatic information, and substantial gaps remain even after end-to-end fine-tuning for downstream tasks. We examine whether downstream adaptations, including linear probing and end-to-end fine-tuning, can be still utilized to recover the lost performance. We find that training with clean and artifact-corrupted images improves performance not only on artifact-corrupted test images but also on clean test images under both linear probing and end-to-end fine-tuning. These improvements, however, transfer only partly to unseen artifacts. Similarly, transfer of robustness gained from artifacts to the color domain, and vice versa, is limited. Therefore, retinal foundation models must be rigorously evaluated to ensure that their performance generalizes to scenarios with possible corruptions observed beyond training. Our results also motivate more principled machine learning solutions to address the broad-spectrum resilience of these models that facilitate diverse ophthalmic research and clinical applications.

1 INTRODUCTION

Retinal foundation models learn representations that can be adapted to many diagnostic tasks using images of the posterior segment of the eye. Color fundus photographs are the most common retinal images. RETFound uses them for self-supervised pretraining (Zhou et al., 2023; 2026), while FLAIR, EyeCLIP, RetiZero, and VisionUnite also use language supervision via Contrastive Language–Image Pretraining (CLIP) (Silva-Rodríguez et al., 2025; Shi et al., 2025; Wang et al., 2025; Li et al., 2025). Their performance reported across diverse downstream tasks positions them as reusable encoders for retinal image analysis. However, clinical adoption would require more than achieving high performance on select images. Models must be robust to changes in imaging quality and conditions in dynamic and heterogeneous healthcare environments.

Fundus photographs can be affected by uneven illumination, movement, and variation in acquisition device and practice (Ray et al., 2005; Bazvand & Ghassemi, 2020; Fu et al., 2019). These changes motivate evaluation beyond clean images. Existing medical robustness benchmarks examine only general-purpose vision-language models (VLMs) and medical CLIP models (Cheng et al., 2025; Imam et al., 2025). We study retinal-specific encoders and ask further questions: i) Are the retinal foundation models that the ophthalmology community is relying on robust to imaging artifacts? ii) How does downstream training affect robustness? iii) Do effects generalize to unseen corruptions?

To answer these questions, we introduce a systematic benchmark of six retinal encoders on 17 binary classification tasks covering five disease categories. We adopt the five artifact categories used by Cheng et al. (2025): bias field, noise, motion, cropping, and rotation. Their benchmark uses intensity changes to represent effects of hardware noise and acquisition procedures, and spatial

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

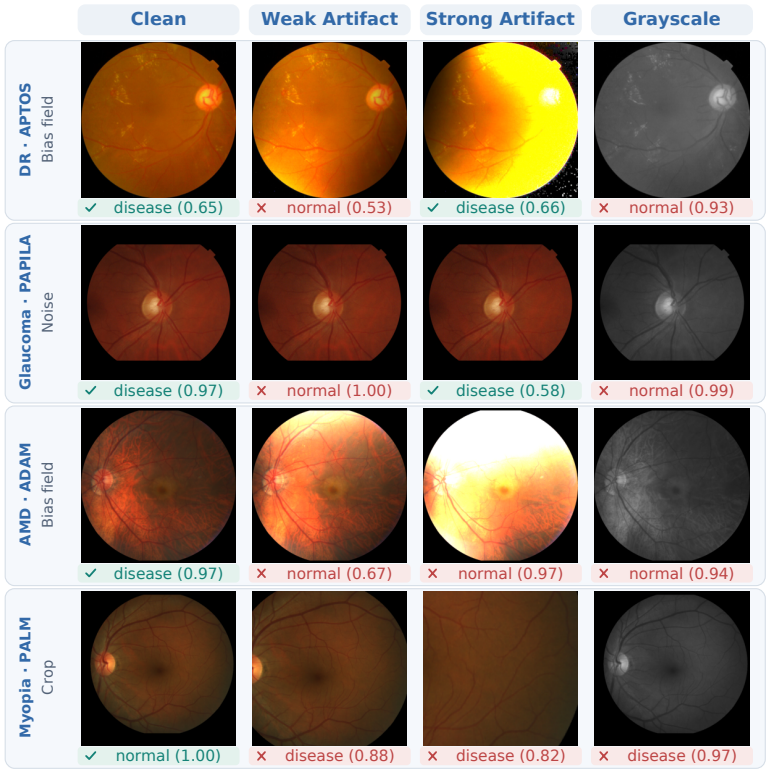

Figure 1: **Examples of prediction changes under synthetic artifacts and color removal.** Rows show held-out images from different disease categories. Columns show the clean image, weak and strong settings of the named artifact, and grayscale conversion. Predictions come from a clean-trained FLAIR-based linear probe; green and red indicate correct and incorrect predictions, with confidence in parentheses. These examples illustrate possible failures, not their frequency or clinical gradability. DR and AMD denote diabetic retinopathy and age-related macular degeneration.

transformations to represent misalignment and a limited field of view. Following this rationale, we test the sensitivity of models to intensity and spatial changes in color fundus images, using two severity settings per artifact (Fig. 1). We also use grayscale as a complementary test of color sensitivity. These controlled artifacts support reproducible comparisons by emulating partially gradable and non-gradable images depending on the artifact severity, as described by Cheng et al. (2025).

Despite their strong overall performance, the models remain highly susceptible to artifacts, with even weak corruptions causing substantial performance degradation. To address this limitation, we investigate whether robustness can be improved through targeted adaptation to corrupted images. We examine how training on clean and corrupted images influences robustness under both linear probing and end-to-end fine-tuning. We also investigate the extent to which robustness learned from certain artifacts transfers to unseen artifact types, as well as whether improvements in artifact robustness extend to robustness against changes in color. Our results show that training with corrupted images can substantially improve robustness even when the image encoder remains frozen; however, robustness transfers only partially across artifact types and from artifact robustness to color sensitivity. While this adaptation-based strategy provides a practical short-term means of improving model robustness, our findings motivate more systematic approaches to robustness that generalize beyond the perturbations encountered during training. We will release our code, corruption-generation procedures, evaluation records, and data manifests with the camera-ready version to support reproducible comparisons.

2 Related Work

RETFound (Zhou et al., 2023) is a suite of two large vision transformers (ViTs) (Dosovitskiy et al., 2021). They were pretrained on 904,170 color fundus images and 736,442 two-dimensional cross-sectional images (B-scans) from optical coherence tomography (OCT) volumes, respectively, via masked image modeling (He et al., 2022). RETFound's representations transfer well to ocular and systemic disease prediction tasks (Zhou et al., 2023). A newer RETFound variant uses the DINOv2 self-distillation objective (Oquab et al., 2024) instead of masked autoencoding (Zhou et al., 2026). We include both variants in our study because their distinct pretraining objectives and augmentation policies can affect robustness. We also consider the vision-language successors such as FLAIR (Silva-Rodríguez et al., 2025), EyeCLIP (Shi et al., 2025), RetiZero (Wang et al., 2025) and VisionUnite (Li et al., 2025). FLAIR is a ResNet–BERT pair trained on 37 assembled fundus datasets with expert-knowledge prompts. EyeCLIP extends CLIP (Radford et al., 2021) to 2.77 million multimodal ophthalmic images. RetiZero adapts RETFound with low-rank adaptation (LoRA) (Hu et al., 2022) to a CLIP embedding space covering 400 fundus diseases. VisionUnite couples image encoders with LLaMA-7B (Touvron et al., 2023) and adapts the whole model to the MMFundus[1] dataset, a large collection of public fundus images paired with descriptive text. However, none of the studies evaluate robustness to imaging artifacts, despite the variety of downstream tasks.

ImageNet-C (Hendrycks & Dietterich, 2019) provides a template for evaluating models on synthetically corrupted natural images. In the medical domain, the work of Cheng et al. (2025) constitutes the closest antecedent to ours: an artifact-robustness benchmark for vision–language models spanning brain MRI, retinal OCT, and chest X-ray images, using five clinically relevant imaging artifact categories, with intensity transformations generated using TorchIO (Pérez-García et al., 2021). They show that classification performance declines under weak artifacts for both proprietary and open VLMs. When severely distorted, ungradable images are used, generative VLMs rarely report their poor image quality. These results are also validated on a public set of color fundus images, treating partially obscured but gradable fundus images and severely distorted, ungradable ones as weak and strong artifact conditions, respectively. In a similar vein, RobustMedCLIP (Imam et al., 2025) studies medical CLIP variants across five imaging modalities and proposes LoRA (Hu et al., 2022) with a few corrupted samples to improve robustness. We adopt the same artifact categories as Cheng et al. (2025) but evaluate six retinal-specific foundation-model encoders using color fundus images in a comprehensive benchmark.

Fundus models often generalize poorly across camera manufacturers, demographic populations, and photographic styles. Color transformation has been shown to reduce the impact of domain changes and photographic style differences and improve performance on unseen fundus data (Xiong et al., 2020; Guo & Ji, 2024). We also use color transformations to understand the robustness of retinal foundation models.

In the newest work, Zoellin et al. (2026) compare 34 encoders on 39 ophthalmic classification tasks using frozen features, linear probing, and end-to-end fine-tuning. They study model selection and adaptation to downstream tasks but do not evaluate robustness to image artifacts. Thus, our benchmark is orthogonal to theirs as we examine robustness to controlled corruptions and their transfer beyond those used during training.

In our benchmark, we use two common methods for evaluating encoders: linear probing and fine-tuning. Linear probing on frozen features is the standard way to compare representations; however, Kumar et al. (2022) showed that fine-tuning can distort pretrained features and lose robustness on unfamiliar data even if accuracy on similar data improves. This motivates our comparison of linear probing and end-to-end fine-tuning.

3 Benchmark Design

We design our benchmark to evaluate i) robustness after clean-data adaptation, ii) recovery through targeted training with artifacts, and iii) transfer of robustness to unseen artifacts. We first compare six retinal encoders on 17 datasets using frozen encoders with linear probes and end-to-end fine-tuning. Training for both linear probing and fine-tuning is initially performed on clean images.

[1] https://github.com/HUANGLIZI/MMFundus

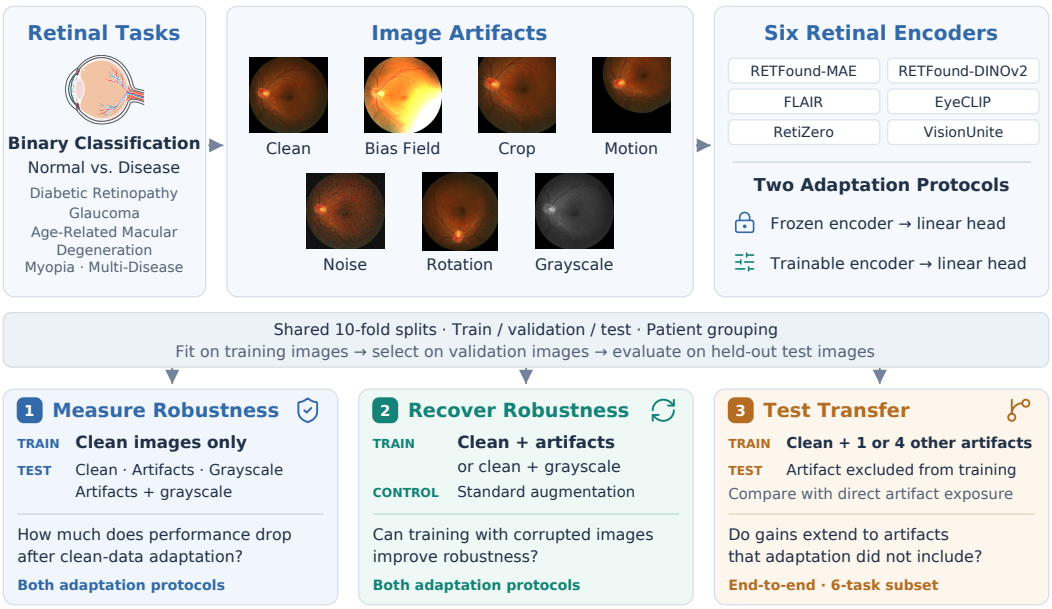

Figure 2: **Benchmark overview.** Retinal classification tasks and six encoders support three analyses: robustness after clean-data adaptation, recovery through corruption training, and transfer of robustness to unseen artifacts. Thumbnails show the five synthetic artifacts, followed by grayscale as a test for color sensitivity. We compare linear probes and end-to-end fine-tuning using shared splits, with patient grouping where necessary. We evaluate cross-artifact transfer and standard end-to-end augmentation control on a subset of six tasks.

We then train models with artifacts and grayscale images to assess the impact of such targeted adaptation on robustness, recovery, and transfer (Sec. 5). We use a six-dataset subset for end-to-end augmentation-control and cross-artifact transfer analyses.

3.1 Models

We evaluate six publicly released retinal encoders: RETFound-MAE, RETFound-DINOv2, FLAIR, EyeCLIP, RetiZero, and VisionUnite. Each retains its native feature dimension, input resolution, and preprocessing. We use the image encoders with task-specific linear classifier heads, comparing frozen representations with end-to-end fine-tuning. Appendix A lists the architectures and input specifications and describes the checkpoints and feature extraction procedures.

3.2 Datasets

We use 17 public datasets covering five disease categories. Due to the diversity of labeling schemes, including multi-class and multi-label scenarios, we formulate each task as binary normal-versus-disease classification by dichotomizing labels. More details about the datasets and our task definitions can be found in Appendix B. We balance each dataset by sampling without replacement $\min(n_{\text{normal}}, n_{\text{disease}}, 500)$ images per class with a fixed seed. This gives a 50% chance baseline for each dataset. We weight datasets equally when reporting aggregate results. Appendix B explains the sampling details.

The resulting benchmark contains 9,728 images: 3,536 from five diabetic retinopathy datasets, 2,964 from five glaucoma datasets, 1,486 from three age-related macular degeneration datasets, 374 from a pathological myopia dataset, and 1,368 from three multi-disease screening datasets. Appendix B describes the label mappings. Several datasets also appear in MMFundus, VisionUnite's pretraining collection (Li et al., 2025). We record this membership in Table 6 and discuss its implications in Sec. 6.

3.3 Artifacts and Test Conditions

We follow Cheng et al. (2025) and use bias field, noise, motion, cropping, and rotation. We generate bias field, noise, and motion artifacts with TorchIO (Pérez-García et al., 2021), and implement cropping and rotation as spatial image transforms. Each artifact has two settings, termed weak and strong, yielding 10 artifact views per image.

These transformations provide repeatable tests of sensitivity to intensity and spatial changes. We do not assess clinical gradability or validate their correspondence to fundus acquisition artifacts. Nevertheless, TorchIO describes its motion transform as a simulation of MRI motion artifacts that combines k-space information from transformed images (TorchIO contributors, n.d.). We therefore use this transform as a synthetic stress test and generate partially gradable and ungradable images, but do not assume that it faithfully reproduces motion degradation in fundus photographs.

To assess color sensitivity alongside artifact robustness, we also convert each clean image to grayscale and apply the same five artifact types at both severities. Grayscale values are replicated across three channels to match the encoders' input format. This produces four test conditions: clean(one view), artifact (ten views), grayscale (one view), and artifact+grayscale (ten views), for 22 views per image. Grayscale-only tests sensitivity to color removal without changing pixel locations.

Before generating these views, we resize each source image to 512x512 pixels while preserving its aspect ratio. This provides a common working scale for datasets with different native resolutions. Random seeds derived from the image identifier and test condition make the views reproducible. All encoders receive the same generated images before their model-specific preprocessing.

3.4 Evaluation Protocol

We use shared 10-fold stratified group cross-validation across encoders and adaptation schemes. Patient identifiers define groups in PAPILA and AMDNet23; elsewhere, each image forms its own group. Training, validation, and test partitions have no groups in common. The area under the receiver operating characteristic curve (AUROC) on the validation set determines checkpoint selection and early stopping. Test data are used only for evaluation.

We train classifiers via both linear probing and end-to-end fine-tuning on clean images, which determines the baselines for all four test conditions (Sec. 4). Frozen-probe features are standardized using clean training data statistics, which also remain fixed at test time. Optimization settings are shared across encoders within each adaptation scheme. Artifact, grayscale, and standard-augmentation experiments retain the splits and corresponding optimization settings, with validation views matched to each intervention (Sec. 5). Appendix C provides the training settings and validation procedures.

To assess robustness, we consider the accuracy on corrupted images and its drop from clean-image accuracy, measured in percentage points. For artifact and artifact+grayscale conditions, we pool predictions across all 10 views per test image. Using weak and strong artifacts together is a deliberate choice for evaluating robustness to a broader range of artifacts, given that strong artifacts do not necessarily lead to worse performance than weak ones; for example, weak rotation reduces accuracy more than strong rotation. Appendix G explains more detail about severity effects.

We average scores across folds and weight datasets equally. We also compute confidence intervals and paired tests from fold-level scores; Appendix C explains these calculations and reports the results.

4 Robustness After Training on Clean Images

We first test whether adapting an encoder to clean diagnostic images provides robustness to artifacts. Table 1 compares linear probing and end-to-end fine-tuning on the same 17 tasks. Both use clean training images.

4.1 All six encoders lose accuracy under artifacts

Linear probing achieves 80.1% mean clean accuracy and 65.8% artifact accuracy, a drop of 14.3 percentage points. Every encoder loses accuracy under the artifact suite. These averages pool weak

Table 1: Main results under linear probing (LP) and end-to-end fine-tuning (FT), trained on clean images with shared data splits. Each entry depicts an average across ten folds and 17 equally weighted datasets. Clean indicates accuracy (%) on clean images; A denotes artifacts and G grayscale; Δ is clean minus corrupted accuracy (percentage points). Artifact pools five types across two severities; $A + G$ combines artifacts with grayscale.

Encoder	Linear Probe				End-to-End Fine-Tuning			
	Clean	Δ_A	Δ_G	Δ_{A+G}	Clean	Δ_A	Δ_G	Δ_{A+G}
RETFound-MAE	77.9	17.5	17.0	23.2	76.3	15.2	13.9	21.1
RETFound-DINOv2	81.0	12.8	**6.5**	**18.5**	**84.7**	11.7	**4.4**	**17.4**
FLAIR	83.0	14.0	16.3	23.1	83.2	12.8	10.6	19.7
EyeCLIP	72.1	**9.5**	18.2	19.7	71.8	**8.0**	16.4	18.8
RetiZero	**83.3**	14.3	21.7	27.2	83.4	10.1	19.7	24.6
VisionUnite	83.0	17.5	28.2	30.6	84.0	17.9	28.3	30.6
Mean	80.1	14.3	18.0	23.7	80.6	12.6	15.5	22.0

and strong settings. Weak artifacts alone reduce mean accuracy by 9.9 points, ranging from 7.0 for RETFound-DINOv2 to 16.1 for RETFound-MAE, and strong artifacts by 18.6 points (Appendix G).

Instead, they reflect the lack of robustness of models to a broader range of artifacts in an aggregate way. Also, a small gap does not necessarily indicate better corrupted-image performance. EyeCLIP has the smallest artifact gap, 9.5 points, but also the lowest clean accuracy, 72.1%. RETFound-DINOv2 combines 81.0% clean accuracy with a 12.8-point gap. FLAIR and RetiZero have higher clean accuracies of 83.0% and 83.3%, with gaps of 14.0 and 14.3 points. Table 14 in Appendix F shows variation across disease categories. We also report results for individual tasks in Appendix E.

4.2 Clean-data fine-tuning leaves substantial gaps

We next allow all encoder parameters to change during adaptation. Fine-tuning can improve task performance while also changing robustness to distribution shift (Kumar et al., 2022). Under our clean-training recipe, mean clean accuracy slightly increases from 80.1% to 80.6%, and artifact accuracy increases from 65.8% to 68.0%. The artifact gap narrows from 14.3 to 12.6 points but remains substantial. The grayscale and combined gaps also remain at 15.5 and 22.0 points.

Overall, the changes differ across encoders and tasks. RETFound-DINOv2 gains 3.7 points in clean accuracy, whereas RETFound-MAE loses 1.6 points. Fine-tuning improves clean accuracy in 58 of 102 encoder–task pairs. Nevertheless, these results show that the observed corruption sensitivity persists beyond linear probing.

4.3 Color removal reveals a different sensitivity profile

Grayscale conversion provides a complementary test because it removes color while preserving pixel locations. With clean-trained linear probes, the mean accuracy gap is 18.0 points, ranging from 6.5 for RETFound-DINOv2 to 28.2 for VisionUnite. Color sensitivity also differs substantially among retinal encoders, and the mean accuracy gap is larger for grayscale than for artifacts.

Combining artifacts and grayscale increases the mean accuracy gap to 23.7 points. Clean-data fine-tuning reduces the grayscale and combined gaps to 15.5 and 22.0 points, respectively, but both remain substantial. Section 5.3 tests whether training with grayscale images reduces this sensitivity and whether the gains extend to artifacts.

5 Robustness Recovery and Transfer

We next test whether training with artifact-corrupted or grayscale images improves robustness, and whether those gains extend to corruptions excluded from training. We also use a standard-augmentation control to assess whether adding more training views explains the gains from artifact training. This control generates the same number of additional views per image using flips, rotation,

Table 2: End-to-end artifact fine-tuning versus a standard-augmentation control on ADAM, APTOS, FIVES, G1020, MMRDR, and PALM. Both augmented regimes add ten views per image and use matching validation views. Results are averaged over ten folds and six equally weighted datasets per encoder. Clean is accuracy (%); A denotes artifacts and G grayscale; Δ is clean minus corrupted accuracy (percentage points).

Encoder	Clean Only			+ Standard Aug.			+ Artifacts		
	Clean	Δ_A	Δ_G	Clean	Δ_A	Δ_G	Clean	Δ_A	Δ_G
RETFound-MAE	78.3	17.4	12.2	84.1	13.8	18.3	85.9	5.4	28.4
RETFound-DINOv2	86.1	11.0	**5.2**	85.4	8.7	**3.3**	88.0	5.4	**2.2**
FLAIR	84.2	11.8	6.9	87.9	9.2	6.5	88.0	4.9	5.2
EyeCLIP	70.1	**7.5**	17.5	79.1	8.4	23.1	81.3	5.4	27.6
RetiZero	83.4	10.9	22.1	85.0	**8.0**	19.6	86.8	**4.1**	28.4
VisionUnite	**91.6**	23.5	37.0	**94.0**	18.0	39.0	**94.9**	14.0	33.9
Mean	82.3	13.7	16.8	85.9	11.0	18.3	87.5	6.5	21.0

resized crops, color jitter, and blur. The primary comparisons cover all six encoders and 17 tasks. The standard-augmentation control and artifact-transfer experiments use six datasets: ADAM, APTOS, FIVES, G1020, MMRDR, and PALM. Appendices D and J explain the additional checks on pretraining augmentations and the evaluation protocol.

5.1 Artifact training improves clean accuracy and robustness

Table 2 compares end-to-end fine-tuning on the six-dataset subset using clean images only, standard augmentation, and artifact augmentation. Both augmented methods add 10 views of each training image and use the same experimental settings. Appendix I reports the complementary linear-probe experiments on all 17 tasks.

Standard augmentation improves mean clean accuracy from 82.3% to 85.9% and reduces the mean artifact gap from 13.7 to 11.0 percentage points. Thus, standard augmentation alone provides some benefit, but substantial accuracy losses remain on corrupted images. Increasing the number of training views with this augmentation pipeline provides only partial robustness to the evaluated artifacts.

Adding artifact-corrupted images improves both clean accuracy and artifact robustness. Mean clean accuracy reaches 87.5%, and the mean artifact gap falls to 6.5 points. Every encoder achieves higher clean accuracy and a smaller artifact gap than with either clean-only training or standard augmentation, when averaged over the six datasets. These improvements suggest that the tested artifact transformations provide useful training variation beyond the standard augmentation pipeline.

However, artifact training does not eliminate sensitivity to image changes. Mean gaps of 6.5 points under artifacts and 21.0 points under grayscale remain, despite the improvement in clean accuracy. Training with these artifact views is therefore insufficient to achieve robustness across the evaluated conditions. The remaining gaps motivate further study of training objectives and methods such as adversarial training.

5.2 Robustness transfers only partly to unseen artifacts

Table 3 compares three end-to-end training methods across all six encoders and the six-dataset subset. Each method uses clean images together with views of either the test artifact (Seen), the other four artifacts (leave-one-artifact-out, LOAO), or one other artifact (One other). The transfer experiments exclude the test artifact from both training and validation. Results average both test severities, and One other also averages the four possible training artifacts.

Training on one other artifact gives a mean test accuracy of 72.4%. Training on all four other artifacts increases this to 75.3%, but accuracy remains below the 79.7% achieved when training includes the test artifact itself. This pattern holds for all five test artifact types when averaged across encoders and datasets. Thus, training on other artifacts provides partial transfer, but does not match the performance of training on the test artifact. LOAO also uses more views per image, so its advantage over One other stems from the differences in both the number and variety of training views.

Table 3: Accuracy (%) on artifact-corrupted test images after end-to-end fine-tuning under three training methods. Seen includes the artifact used at test time; LOAO excludes that artifact and includes the other four; One other uses a single different artifact, with results averaged over the four possible choices. Counts below the column headings include the clean image and its augmented views. Results average six encoders, the six datasets listed in Table 2, ten folds, and both test severities. Gap is the accuracy difference between Seen and LOAO, in percentage points. Bold marks the highest accuracy within each row.

Test artifact	Encoder trained on clean images plus …			Gap
	Seen (3 views)	LOAO (9 views)	One other (3 views)	
Bias field	82.6	78.6	76.3	4.0
Crop	79.9	74.6	70.8	5.3
Motion	72.5	72.0	68.9	**0.5**
Noise	80.1	71.0	69.0	9.1
Rotation	**83.4**	**80.5**	**77.0**	2.9
Mean	79.7	75.3	72.4	4.4

Table 4: End-to-end fine-tuning with clean images, clean plus grayscale, or clean plus ten artifact views. Training and validation use the same view composition. Results average ten folds and 17 equally weighted datasets per encoder. Clean is accuracy (%); A denotes artifacts and G grayscale; Δ is clean minus corrupted accuracy (percentage points). Bold marks the highest clean accuracy or smallest gap in each column across encoders. A larger gap need not imply lower corrupted accuracy when clean accuracy also changes.

Encoder	Clean Only			+ Grayscale			+ Artifacts		
	Clean	Δ_A	Δ_G	Clean	Δ_A	Δ_G	Clean	Δ_A	Δ_G
RETFound-MAE	76.3	15.2	13.9	77.6	16.2	1.0	84.5	4.5	23.4
RETFound-DINOv2	**84.7**	11.7	**4.4**	85.3	11.8	**0.3**	86.7	4.8	**3.6**
FLAIR	83.2	12.8	10.6	85.5	13.5	1.5	87.3	5.5	7.6
EyeCLIP	71.8	**8.0**	16.4	73.0	**8.6**	2.5	79.4	4.7	21.5
RetiZero	83.4	10.1	19.7	84.2	10.1	1.9	85.6	**4.0**	21.6
VisionUnite	84.0	17.9	28.3	**85.7**	18.2	8.7	**87.5**	10.4	28.5
Mean	80.6	12.6	15.5	81.9	13.1	2.7	85.2	5.6	17.7

The transfer gap differs across artifact types. Excluding noise from training reduces accuracy by 9.1 percentage points compared with Seen, followed by a 5.3-point reduction for cropping. These artifacts benefit most from direct exposure during training. Motion has the smallest gap at 0.5 points, but its accuracy is only 72.5% even when motion is included in training. A small transfer gap therefore does not necessarily indicate strong performance on the artifact.

Together, these results show that gains on artifacts included during training do not fully extend to unseen artifact types. Robustness evaluation should therefore include artifacts excluded from training. Appendix H explains transfer between individual artifact types and reports the full comparison in Table 16.

5.3 Artifact and grayscale training have different effects

Table 4 tests whether improvements in artifact robustness also reduce sensitivity to color removal, and whether the reverse holds. We compare end-to-end fine-tuning on clean images only, clean plus grayscale images, and clean plus 10 artifact views per image, using the same experimental settings. Here, grayscale training replaces the standard-augmentation control in Table 2, and the comparison covers all six encoders and all 17 tasks. Appendix I reports the complementary linear-probe experiments.

Grayscale training reduces the mean grayscale gap from 15.5 to 2.7 percentage points and improves mean clean accuracy from 80.6% to 81.9%. However, the mean artifact gap remains at 13.1 points,

compared with 12.6 points under clean-only training. Thus, reducing sensitivity to color removal does not reduce the average accuracy loss under artifacts.

Artifact training reduces the mean artifact gap from 12.6 to 5.6 points, and increases mean clean accuracy to 85.2%. Every encoder achieves higher clean accuracy than with either clean-only or grayscale training, when averaged over the 17 tasks. However, the mean grayscale gap increases to 17.7 points. Four encoders have larger grayscale gaps after artifact training; RETFound-DINOv2 and FLAIR are the exceptions. Artifact training therefore improves clean performance and artifact robustness, but does not consistently reduce color sensitivity.

The gaps must also be interpreted alongside absolute accuracy. Artifact training raises mean grayscale accuracy from 65.1% to 67.5%, even though the grayscale gap grows, because clean accuracy improves more. The larger gap therefore does not indicate lower grayscale accuracy. Together, these results support evaluating artifact robustness and color sensitivity as two separate axes: training on one type of image change does not consistently address the other.

6 Discussion

Foundation models have driven unprecedented advances in medical image analysis and disease diagnosis. In this study, we focus on retinal foundation models trained on color fundus images and systematically investigate their robustness to imaging artifacts under two main evaluation protocols: linear probing and end-to-end fine-tuning. Interestingly, these models the ophthalmology community is building on are susceptible to simple imaging artifacts, despite their overall performance across diverse diagnostic tasks. As a remedy, we explore whether their performance could be recovered through targeted adaptations to artifacts.

Interestingly, robustness can be improved with targeted adaptations using artifact-corrupted images under both linear probing and end-to-end fine-tuning. Targeted training with artifacts also leads to higher clean accuracy and a smaller artifact gap than with standard data augmentation. These results highlight the value of considering potential imaging artifacts during downstream task adaptation. On top of that, the gains with frozen encoders show that existing pretrained features support robust classification through similar changes to the linear probe training.

These improvements partially transfer between artifact types, with the benefit of direct exposure varying across artifact types. Artifact and grayscale training also produce different effects, supporting their evaluation as two separate robustness axes. These findings motivate testing joint artifact and grayscale training objectives that encourage transfer across image changes.

Some limitations of this work suggest directions for further study. The experiments use synthetic artifacts and balanced binary tasks. Evaluation on naturally degraded images, with expert assessment of diagnostic information, can help connect these findings to clinical practice. More complete patient identifiers and checks for image overlap with pretraining datasets, including MMFundus, can also strengthen the evaluation. Similarly, well-calibrated predictive uncertainty estimates could enable models to abstain from making predictions under artifact-induced ambiguity by reflecting the inherent difficulties that retinal specialists face when grading fundus images (Ayhan et al., 2020).

A future direction is to develop more principled approaches to the pretraining or adaptation of foundation models that will address both artifact robustness and color sensitivity, and test whether the gains extend to unseen image changes.

7 Conclusion

We benchmark six retinal foundation models on 17 diagnostic tasks to examine their robustness to synthetic artifacts and color removal. All six models exhibit substantial performance gaps on artifacts after training on clean images. Training with artifact-corrupted images improves performance on both clean and corrupted images under both linear probing and end-to-end fine-tuning. These improvements, however, transfer only partly to unseen artifacts and do not consistently extend between artifact robustness and color sensitivity. Our benchmark provides a shared setting for these comparisons and for evaluating potential training methods that address both robustness axes.

REFERENCES

Jin Mo Ahn, Sangsoo Kim, Kwang-Sung Ahn, Sung-Hoon Cho, Kwan Bok Lee, and Ungsoo Samuel Kim. A deep learning model for the detection of both advanced and early glaucoma using fundus photography. *PLOS ONE*, 13(11):e0207982, 2018. doi: 10.1371/journal.pone.0207982.

Md. Aiyub Ali, Md. Shakhawat Hossain, Md. Kawsar Hossain, Subhadra Soumi Sikder, Sharun Akter Khushbu, and Mirajul Islam. AMDNet23: Hybrid CNN-LSTM deep learning approach with enhanced preprocessing for age-related macular degeneration (AMD) detection. *Intelligent Systems with Applications*, 21:200334, 2024. doi: 10.1016/j.iswa.2024.200334.

Asia Pacific Tele-Ophthalmology Society. APTOS 2019 blindness detection. Kaggle competition, 2019. URL https://www.kaggle.com/competitions/aptos2019-blindness-detection.

Murat Seçkin Ayhan, Laura Kühlewein, Gulnar Aliyeva, Werner Inhoffen, Focke Ziemssen, and Philipp Berens. Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. *Medical Image Analysis*, 64:101724, 2020. doi: 10.1016/j.media.2020.101724.

Muhammad Naseer Bajwa, Gur Amrit Pal Singh, Wolfgang Neumeier, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection. In *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, 2020. doi: 10.1109/IJCNN48605.2020.9207664.

Fatemeh Bazvand and Fariba Ghassemi. Artifacts in macular optical coherence tomography. *Journal of Current Ophthalmology*, 32(2):123–131, 2020. doi: 10.4103/joco.joco_83_20.

Ling-Ping Cen, Jie Ji, Jian-Wei Lin, Si-Tong Ju, Hong-Jie Lin, Tai-Ping Li, Yun Wang, Jian-Feng Yang, Yu-Fen Liu, Shaoying Tan, et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nature Communications*, 12:4828, 2021. doi: 10.1038/s41467-021-25138-w.

Zijie Cheng, Ariel Yuhan Ong, Siegfried K. Wagner, David A. Merle, Lie Ju, Hanyuan Zhang, Ruinian Chen, Linze Pang, Boxuan Li, Tiantian He, Anran Ran, Hongyang Jiang, Dawei Gabriel Yang, Ke Zou, Jocelyn Hui Lin Goh, Sahana Srinivasan, Andre Altmann, Daniel C. Alexander, Carol Y. Cheung, Yih Chung Tham, Pearse A. Keane, and Yukun Zhou. Understanding the robustness of vision-language models to medical image artefacts. *npj Digital Medicine*, 8:727, 2025. doi: 10.1038/s41746-025-02108-w.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Huihui Fang, Fei Li, Huazhu Fu, Xu Sun, Xingxing Cao, Fengbin Lin, Jaemin Son, Sunho Kim, Gwenolé Quellec, et al. ADAM challenge: Detecting age-related macular degeneration from fundus images. *IEEE Transactions on Medical Imaging*, 41(10):2828–2847, 2022. doi: 10.1109/TMI.2022.3172773.

Huihui Fang, Fei Li, Junde Wu, Huazhu Fu, Xu Sun, José Ignacio Orlando, Hrvoje Bogunović, Xiulan Zhang, and Yanwu Xu. Open fundus photograph dataset with pathologic myopia recognition and anatomical structure annotation. *Scientific Data*, 11:99, 2024a. doi: 10.1038/s41597-024-02911-2.

Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024b. doi: 10.1016/j.imavis.2024.105171.

Huazhu Fu, Boyang Wang, Jianbing Shen, Shanshan Cui, Yanwu Xu, Jiang Liu, and Ling Shao. Evaluation of retinal image quality assessment networks in different color-spaces. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Lecture Notes in Computer Science, pp. 48–56. Springer, 2019. doi: 10.1007/978-3-030-32239-7_6.

Song Guo and Ke Ji. Random color transformation for single domain generalized retinal image segmentation. *Engineering Applications of Artificial Intelligence*, 136:108907, 2024. doi: 10.1016/j.engappai.2024.108907.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Raza Imam, Rufael Marew, and Mohammad Yaqub. On the robustness of medical vision-language models: Are they truly generalizable? In *Medical Image Understanding and Analysis (MIUA)*, Lecture Notes in Computer Science, pp. 233–256. Springer, 2025. doi: 10.1007/978-3-031-98688-8_17.

Kai Jin, Xingru Huang, Jingxing Zhou, Yunxiang Li, Yan Yan, Yibao Sun, Qianni Zhang, Yaqi Wang, and Juan Ye. FIVES: A fundus image dataset for artificial intelligence based vessel segmentation. *Scientific Data*, 9:475, 2022. doi: 10.1038/s41597-022-01564-3.

Ungsoo Kim. Machine learn for glaucoma. Harvard Dataverse, 2018. URL https://doi.org/10.7910/DVN/1YRRAC.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL https://arxiv.org/abs/1412.6980.

Oleksandr Kovalyk, Juan Morales-Sánchez, Rafael Verdú-Monedero, Inmaculada Sendán-Alcaraz, and José-Luis Sancho-Gómez. PAPILA: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. *Scientific Data*, 9:291, 2022. doi: 10.1038/s41597-022-01388-1.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022.

Zihan Li, Diping Song, Zefeng Yang, Deming Wang, Fei Li, Xiulan Zhang, Paul E. Kinahan, and Yu Qiao. VisionUnite: A vision-language foundation model for ophthalmology enhanced with clinical knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12): 11848–11862, 2025. doi: 10.1109/TPAMI.2025.3598734.

Chi Liu, Xiaotong Han, Zhixi Li, Jason Ha, Guankai Peng, Wei Meng, and Mingguang He. A self-adaptive deep learning method for automated eye laterality detection based on color fundus photography. *PLOS ONE*, 14(9):e0222025, 2019. doi: 10.1371/journal.pone.0222025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick

Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt.

Samiksha Pachade, Prasanna Porwal, Dhanshree Thulkar, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, Luca Giancardo, Gwenolé Quellec, and Fabrice Mériaudeau. Retinal fundus multi-disease image dataset (RFMiD): A dataset for multi-disease detection research. *Data*, 6(2): 14, 2021. doi: 10.3390/data6020014.

Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin. TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine*, 208:106236, 2021. doi: 10.1016/j.cmpb.2021.106236.

Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (IDRiD): A database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018. doi: 10.3390/data3030025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Reena Ray, Sandra S. Stinnett, and Glenn J. Jaffe. Evaluation of image artifact produced by optical coherence tomography of retinal pathology. *American Journal of Ophthalmology*, 139(1):18–29, 2005. doi: 10.1016/j.ajo.2004.07.050.

Shayla Sharmin, Mohammad Riadur Rashid, Tania Khatun, Md Zahid Hasan, Mohammad Shorif Uddin, and Marzia. A dataset of color fundus images for the detection and classification of eye diseases. *Data in Brief*, 57:110979, 2024. doi: 10.1016/j.dib.2024.110979.

Danli Shi, Weiyi Zhang, Jiancheng Yang, Siyu Huang, Xiaolan Chen, et al. A multimodal visual–language foundation model for computational ophthalmology. *npj Digital Medicine*, 8:381, 2025. doi: 10.1038/s41746-025-01772-2.

Julio Silva-Rodríguez, Hadi Chakor, Riadh Kobbi, Jose Dolz, and Ismail Ben Ayed. A foundation language-image model of the retina (FLAIR): encoding expert knowledge in text supervision. *Medical Image Analysis*, 99:103357, 2025. doi: 10.1016/j.media.2024.103357.

Zhenyu Tang, Lilong Wang, Zhen Guo, Qianqian Liang, Shuyue Xue, Chengcheng Feng, Lili Ran, Lingzhi Chen, Xiaosong Wang, and Jun Li. A multimodal retinal image dataset for diabetic retinopathy detection using foundation models. *Scientific Data*, 13:639, 2026. doi: 10.1038/s41597-026-07005-9.

TorchIO contributors. TorchIO documentation: Motion. https://docs.torchio.org/2.0/reference/transforms/motion/, n.d. Accessed September 8, 2026.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Meng Wang, Tian Lin, Aidi Lin, Kai Yu, Yuanyuan Peng, Lianyu Wang, Cheng Chen, Ke Zou, Huiyu Liang, Man Chen, Xue Yao, Meiqin Zhang, Binwei Huang, Chaoxin Zheng, et al. Enhancing diagnostic accuracy in rare and common fundus diseases with a knowledge-rich vision-language model. *Nature Communications*, 16:5528, 2025. doi: 10.1038/s41467-025-60577-9.

Xue Xia, Ying Li, Guobei Xiao, Kun Zhan, Jinhua Yan, Chao Cai, Yuming Fang, and Guofu Huang. Benchmarking deep models on retinal fundus disease diagnosis and a large-scale dataset. *Signal Processing: Image Communication*, 127:117151, 2024. doi: 10.1016/j.image.2024.117151.

Jianhao Xiong, Andre Wang He, Meng Fu, Xinyue Hu, Yifan Zhang, Congxin Liu, Xin Zhao, and Zongyuan Ge. Improve unseen domain generalization via enhanced local color transformation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 433–443. Springer, 2020. doi: 10.1007/978-3-030-59713-9_42.

Yukun Zhou, Mark A. Chia, Siegfried K. Wagner, Murat S. Ayhan, Dominic J. Williamson, Robbert R. Struyven, Timing Liu, Moucheng Xu, Mateo G. Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981): 156–163, 2023. doi: 10.1038/s41586-023-06555-x.

Yukun Zhou, Zheyuan Wang, Yilan Wu, Ariel Yuhan Ong, Siegfried K. Wagner, et al. Understanding pre-training data effects in retinal foundation models using two large fundus cohorts. *Nature Communications*, 17:3309, 2026. doi: 10.1038/s41467-026-70077-z.

Jay Zoellin, Colin Merk, and Bence György. Benchmarking vision encoders for image classification in ophthalmology. *Computational and Structural Biotechnology Journal*, 35(2):0178, 2026. doi: 10.34133/csbj.0178.

A Benchmarked Models

Table 5 summarizes the six encoders and the image representations used in our benchmark. We evaluate their visual features under the shared adaptation protocols, without using text prompts or language-model outputs.

Table 5: Retinal encoders used in the benchmark. D is the dimension of the image representation used for classification. Input resolution denotes the preprocessed image size in pixels; VisionUnite internally resizes images to 224×224 for its CLIP branch.

Model	Vision encoder	D	Input resolution
RETFound-MAE (Zhou et al., 2023)	ViT-L/16 (MAE)	1024	224×224
RETFound-DINOv2 (Zhou et al., 2026)	ViT-L/14 (DINOv2)	1024	224×224
FLAIR (Silva-Rodríguez et al., 2025)	ResNet-50	512	512×512
EyeCLIP (Shi et al., 2025)	CLIP ViT-B/32	512	224×224
RetiZero (Wang et al., 2025)	ViT-L/16 (LoRA)	512	224×224
VisionUnite (Li et al., 2025)	EVA-02 + CLIP ViT-L/14	768	448×448

RETFound-MAE uses masked autoencoding (He et al., 2022); we load its color fundus photography checkpoint and extract the `CLS` representation. For RETFound-DINOv2, we use the released teacher checkpoint trained on Moorfields Eye Hospital (MEH) data. Both encoders provide 1024-dimensional features.

FLAIR and EyeCLIP provide 512-dimensional image embeddings from a ResNet-50 and a CLIP-based ViT-B/32 (Radford et al., 2021), respectively. RetiZero uses a LoRA-adapted (Hu et al., 2022) RETFound-MAE encoder with features projected into a 512-dimensional space. VisionUnite combines EVA-02 (Fang et al., 2024b) and CLIP ViT-L/14 visual features by addition to produce the 768-dimensional representation used in our experiments.

Preprocessing remains model-specific. For frozen feature extraction, we call the released FLAIR and EyeCLIP preprocessing modules directly and implement the pipelines for the other four encoders in our benchmark wrappers. The fine-tuning wrappers implement the corresponding preprocessing for batched training. Although FLAIR and RetiZero share code, their image preprocessing differs: FLAIR scales intensities to $[0, 1]$ and resizes and pads images to 512×512, whereas RetiZero resizes images to 224×224 and applies ImageNet normalization. Both models return ℓ_2-normalized image embeddings.

B Benchmark Datasets

Table 6 summarizes the 17 binary classification tasks across five disease categories. We balance each task by sampling the same number of images from each class, capped at 500 per class, using a fixed

random seed of 42. For APTOS, G1020, Harvard, FIVES, and ADAM, we sample from an earlier pool drawn from the MMFundus copies of these datasets, with labels taken from the MMFundus annotation files. This pool held 100 images per class for APTOS, 80 for ADAM, and fewer images than the full datasets for G1020, Harvard, and FIVES. The remaining twelve tasks sample directly from the public releases. The resulting tasks contain 9,728 image entries in total. EDDFS contributes separate tasks for diabetic retinopathy, glaucoma, and age-related macular degeneration. This total sums the task sizes rather than counting unique images across tasks.

Table 6: Datasets and balanced sample counts for the 17 classification tasks. Each task contains n normal and n positive images. MMF records whether the dataset is included in MMFundus, used to train VisionUnite; dataset membership alone does not establish image-level overlap with pretraining.

Dataset	Source / description	n	MMF
Diabetic retinopathy			
MMRDR (Tang et al., 2026)	China (multi-modal, 2024)	500	no
APTOS (Asia Pacific Tele-Ophthalmology Society, 2019)	India (Aravind)	100	yes
IDRiD (Porwal et al., 2018)	India	168	yes
EDDFS (Xia et al., 2024)	China (multi-label)	500	no
Yangxi (Liu et al., 2019)	China (gray background)	500	no
Glaucoma			
AnwaraHamida (Sharmin et al., 2024)	Bangladesh	500	no
EDDFS (Xia et al., 2024)	China	500	no
G1020 (Bajwa et al., 2020)	Germany	168	yes
Harvard (Kim, 2018; Ahn et al., 2018)	South Korea	159	yes
PAPILA (Kovalyk et al., 2022)	Spain (both eyes)	155	yes
Age-related macular degeneration			
ADAM (Fang et al., 2022)	MICCAI challenge	80	yes
AMDNet23 (Ali et al., 2024)	folder-labeled	163	no
EDDFS (Xia et al., 2024)	China	500	no
Pathological myopia			
PALM (Fang et al., 2024a)	PALM challenge	187	yes
Multi-disease			
FIVES (Jin et al., 2022)	Vessel segmentation dataset	146	yes
JSIEC (Cen et al., 2021)	39-class release	38	yes
RFMiD (Pachade et al., 2021)	India (45 labels)	500	yes

We map source annotations to normal and positive classes for each task. For diabetic retinopathy grading datasets, grade 0 defines the normal class and higher disease grades define the positive class. For EDDFS, each task compares images marked normal with images positive for the target disease; a label for another disease alone does not qualify an image for that task. AnwaraHamida and AMDNet23 use the healthy or normal folder and the target disease folder. For PAPILA, we combine glaucoma and suspected glaucoma into the positive class. The multi-disease tasks combine the selected non-normal categories, with RFMiD using its binary Disease_Risk label. These mappings define the binary tasks used throughout the benchmark.

C Training and Evaluation Details

This section provides the full training, validation, and aggregation procedures for the evaluation protocol in Sec. 3.4.

We use 10-fold stratified group cross-validation, with shared splits across encoders and evaluation schemes. Patient identifiers derived from filenames define groups in PAPILA and AMDNet23; each image forms its own group in the remaining datasets. For each fold, we hold out one partition for testing and reserve 15% of the remaining groups for validation. Training, validation, and test partitions do not share groups. The validation set determines early stopping and checkpoint selection; the test set is used only for evaluation.

For the frozen-probe baseline, we attach a linear classifier $\mathrm{Linear}(D \to 2)$ to each encoder, retaining its native feature dimension $D \in \{512, 768, 1024\}$. We standardize each feature dimension using the mean and standard deviation of the clean training features. We optimize cross-entropy with Adam (Kingma & Ba, 2015), using a learning rate of 10^{-3}, a weight-decay parameter of 10^{-4}, and a batch size of 256. Training runs for at most 200 epochs, with early stopping after 20 epochs without an improvement in validation AUROC. We restore the best validation checkpoint for evaluation. This linear probe-training recipe is fixed across encoders.

We fit each baseline probe once on clean training features and reuse it across all four test conditions. Feature-standardization statistics also remain fixed, so corrupted test data cannot alter the classifier or normalization. This protocol measures robustness without adaptation to the test corruption. We separately study adaptation through artifact and grayscale training in Sec. 5.

For the end-to-end baseline, we initialize each fold from the published encoder checkpoint and train it with a new linear classification head and a learnable logit scale. As in the frozen-probe baseline, training uses clean images without additional data augmentation, and evaluation covers all four test conditions. We optimize cross-entropy with AdamW (Loshchilov & Hutter, 2019), using a batch size of 32, encoder and head learning rates of 10^{-5} and 10^{-3}, respectively, and weight decay of 0.05. Training runs for at most 20 epochs, with two warm-up epochs followed by cosine learning-rate decay. We select the checkpoint with the highest clean validation AUROC and stop after five epochs without improvement. Thus, the frozen-probe and end-to-end baselines use Adam and AdamW, respectively, with a fixed recipe across encoders within each scheme.

The augmentation experiments extend these baselines by adding artifact, grayscale, or standard augmentation views to the training data. They retain the data partitions and the corresponding optimization settings, but use validation views appropriate to each intervention. Both protocols use validation views in the same proportions as their corresponding training data. Section 5 details the training methods, matched controls, and held-out-artifact evaluations.

We report accuracy for each encoder, dataset, and test condition and use validation AUROC for checkpoint selection. For artifact and artifact+grayscale conditions, we pool predictions across all 10 views per test image. We average scores across the ten folds and give each dataset equal weight in aggregate comparisons.

For each encoder, dataset, and test condition, we compute the mean of the ten fold-level scores and a two-sided 95% confidence interval based on the t-distribution, which the per-dataset tables in Appendix E report. Because all encoders and training methods share the same splits, we compare training methods with two-sided paired t-tests over aligned (dataset, fold) pairs within each disease category. We apply these tests to end-to-end fine-tuning versus the linear probe, and to linear probes trained with artifact and grayscale views versus clean-trained probes. Table 7 reports the mean differences and p-values.

D Pretraining Augmentation Audit

Table 8 summarizes the available information on pretraining augmentation for the six encoders. We use this comparison to provide context for their different responses to color removal.

The released FLAIR and RetiZero code uses horizontal flips, small affine transformations, and mild brightness and contrast jitter. The RETFound-DINOv2 publication reports a broader augmentation pipeline, including color jitter, blur, and solarization (Zhou et al., 2026). Its description does not explicitly list grayscale conversion, although the standard DINOv2 implementation includes it. We therefore do not infer that grayscale was absent from pretraining. Similarly, deterministic preprocessing in released inference code does not establish the augmentation used during training.

RETFound-DINOv2 has the smallest mean grayscale accuracy gap under clean-trained linear probing, at 6.5 percentage points (Table 1). Its reported color augmentation motivates further study of how pretraining affects color sensitivity. However, the encoders also differ in architecture, training data, and objectives, so this comparison does not isolate the effect of augmentation. Section 5.3 examines the effects of artifact and grayscale training within each model during downstream adaptation.

Table 7: Mean differences (percentage points) and two-sided paired t-test p-values for each encoder and disease category, computed over aligned (dataset, fold) pairs. *FT vs. LP* compares end-to-end fine-tuning with the frozen linear probe, both trained on clean images: Clean is the change in clean accuracy (FT minus LP), and Δ_A, Δ_G, and Δ_{A+G} are the changes in the clean-to-corrupted accuracy drop (negative values mean fine-tuning narrows the drop). *Artifact LP* and *Grayscale LP* give the accuracy gain on artifact and grayscale test images, respectively, of linear probes trained with the corresponding views over clean-trained probes. Myopia contains one dataset (PALM). Because training sets overlap across folds, p-values are approximate.

	FT vs. LP								Artifact LP		Grayscale LP	
	Clean		Δ_A		Δ_G		Δ_{A+G}		Artifact acc.		Grayscale acc.	
Category	Diff.	p	Diff.	p	Diff.	p	Diff.	p	Gain	p	Gain	p
RETFound-MAE												
DR	−0.2	0.80	−4.8	<0.001	−6.2	0.004	−3.9	<0.001	+17.8	<0.001	+19.8	<0.001
Glaucoma	−1.2	0.33	−0.9	0.43	−2.6	0.09	−1.2	0.36	+9.7	<0.001	+13.3	<0.001
AMD	−3.9	0.02	−3.7	0.02	+4.8	0.09	−1.1	0.59	+16.1	<0.001	+9.1	<0.001
Myopia	−6.4	0.22	−3.0	0.41	+8.5	0.02	−5.0	0.13	+26.7	<0.001	+9.9	<0.001
Multi-disease	−0.6	0.81	+0.9	0.69	−10.3	0.01	−0.5	0.83	+17.4	<0.001	+24.0	<0.001
RETFound-DINOv2												
DR	+5.8	<0.001	−2.4	<0.001	−3.4	0.001	−1.9	0.01	+9.9	<0.001	+6.2	<0.001
Glaucoma	+2.1	0.09	−2.2	0.04	−1.5	0.29	−1.3	0.22	+7.2	<0.001	+5.9	<0.001
AMD	+2.7	0.04	−0.5	0.69	−2.5	0.09	+0.4	0.76	+9.5	<0.001	+8.3	<0.001
Myopia	+2.1	0.19	−1.9	0.24	+3.7	0.14	−5.7	0.010	+10.7	<0.001	+2.7	0.04
Multi-disease	+4.6	0.04	+2.3	0.22	−2.1	0.34	+0.7	0.71	+8.0	<0.001	+5.5	0.03
FLAIR												
DR	+2.1	0.30	−0.2	0.88	−7.8	0.004	−3.9	0.06	+8.8	<0.001	+20.9	<0.001
Glaucoma	−1.0	0.43	−2.7	0.02	+0.5	0.78	−0.2	0.88	+6.5	<0.001	+8.2	<0.001
AMD	+0.8	0.52	+1.1	0.36	+1.6	0.42	+2.1	0.13	+7.3	<0.001	+5.7	<0.001
Myopia	+2.9	0.03	−2.1	0.08	−36.7	<0.001	−24.1	<0.001	+9.9	<0.001	+39.4	<0.001
Multi-disease	−2.6	0.16	−2.5	0.06	−9.7	0.002	−6.6	0.003	+10.0	<0.001	+18.1	<0.001
EyeCLIP												
DR	+0.9	0.47	−1.8	0.02	−3.8	0.009	−0.2	0.81	+9.2	<0.001	+20.6	<0.001
Glaucoma	+2.0	0.06	−1.2	0.20	+0.3	0.84	+1.0	0.39	+6.0	<0.001	+14.0	<0.001
AMD	−2.6	0.09	−2.7	0.010	−1.5	0.54	−3.0	0.11	+6.7	<0.001	+13.9	<0.001
Myopia	−1.1	0.49	+1.7	0.17	−1.1	0.49	−1.1	0.48	+6.5	<0.001	+35.8	<0.001
Multi-disease	−4.0	0.09	−1.2	0.54	−2.6	0.39	−2.9	0.33	+7.1	<0.001	+10.4	<0.001
RetiZero												
DR	+1.7	0.03	−4.2	<0.001	−7.9	<0.001	−2.5	0.03	+11.6	<0.001	+30.9	<0.001
Glaucoma	+1.5	0.08	−4.4	<0.001	−3.1	0.05	−2.6	0.01	+8.9	<0.001	+15.6	<0.001
AMD	−0.4	0.78	−0.8	0.59	+0.2	0.94	−1.4	0.49	+8.0	<0.001	+14.0	<0.001
Myopia	−5.0	0.27	−4.7	0.07	+10.1	0.10	−6.5	0.09	+13.0	<0.001	+7.5	<0.001
Multi-disease	−3.0	0.18	−7.0	0.003	+3.7	0.25	−2.7	0.33	+13.9	<0.001	+21.5	<0.001
VisionUnite												
DR	+3.9	0.01	+1.6	0.21	+4.9	0.008	+4.5	0.007	+8.2	<0.001	+19.4	<0.001
Glaucoma	+1.9	0.07	+0.9	0.29	+2.5	0.09	+0.9	0.36	+4.7	<0.001	+9.5	<0.001
AMD	+0.5	0.81	+0.5	0.78	−4.7	0.06	−1.7	0.37	+10.6	<0.001	+18.9	<0.001
Myopia	−6.5	0.21	−1.2	0.63	−14.6	0.08	−15.8	0.02	+8.1	<0.001	+44.1	<0.001
Multi-disease	−2.5	0.33	−1.9	0.37	−2.5	0.44	−2.3	0.41	+7.4	<0.001	+13.1	<0.001

Table 8: Pretraining transforms reported in publications or identified in released code. Crop ranges denote the fraction of image area retained, and p denotes application probability. The table summarizes available evidence rather than a complete record of every checkpoint's training. EyeCLIP's entry describes its released preprocessing pipeline.

Model	Pretraining augmentation
RETFound-MAE	Random resized crop (0.2–1.0), horizontal flip
RETFound-DINOv2	Global and local crops, horizontal flip, color jitter, Gaussian blur, solarization
FLAIR	Horizontal flip, affine ($\pm 5^\circ$, scale 0.9–1.0, p=0.25), color jitter (brightness/contrast 0.2, p=0.25)
RetiZero	Same transforms as FLAIR in released code
EyeCLIP	Resize, center crop, normalization; pretraining augmentation not established from this pipeline
VisionUnite	Random resized crop (0.9–1.0)

E PER-DATASET RESULTS

Tables 9–13 report accuracy for all six encoders on the 17 tasks, grouped by disease category. Each encoder remains frozen, and its linear probe is trained on clean images and evaluated under all four test conditions. These results provide the task-level comparisons behind the aggregate results in Sec. 4. Reporting absolute accuracy for each condition shows both clean performance and the accuracy retained after image degradation.

Table 9: Accuracy (%) on diabetic retinopathy tasks with frozen encoders and clean-trained linear probes. Values are means over ten folds ± the half-width of the 95% t-based confidence interval. Artifact conditions pool predictions across five artifact types at two severities; A+G combines these artifacts with grayscale. Bold marks the highest accuracy across encoders for each dataset and test condition, using unrounded values.

Dataset	Encoder	Clean	Artifact	Grayscale	A+G
MMRDR_DR	RETFound-MAE	77.4±2.4	59.2±1.0	67.0±2.7	55.6±1.6
	RETFound-DINOv2	82.5±1.9	72.9±1.4	81.1±1.7	**68.2**±1.1
	FLAIR	**93.1**±1.1	**80.1**±1.4	**81.4**±3.5	68.0±1.3
	EyeCLIP	69.8±2.3	63.4±1.8	56.4±3.6	54.4±1.7
	RetiZero	85.5±2.8	74.3±1.8	66.2±2.9	62.1±1.8
	VisionUnite	79.6±3.1	68.1±1.8	64.3±3.3	60.1±2.1
APTOS3662	RETFound-MAE	91.0±5.3	68.5±3.7	71.5±8.4	58.1±3.6
	RETFound-DINOv2	94.0±2.8	**78.8**±3.4	**86.0**±5.5	**71.8**±5.7
	FLAIR	**96.5**±2.9	70.8±4.4	54.0±3.3	52.7±0.7
	EyeCLIP	82.0±6.4	63.9±3.6	51.5±4.5	51.4±1.8
	RetiZero	95.5±4.6	74.5±2.8	50.0±0.0	50.5±1.5
	VisionUnite	95.5±3.1	73.6±3.7	50.0±0.0	52.4±1.5
IDRiD	RETFound-MAE	79.2±6.0	58.9±2.1	53.8±3.3	53.5±1.1
	RETFound-DINOv2	84.5±4.7	69.6±2.9	**78.9**±5.3	**65.8**±3.0
	FLAIR	**90.8**±4.9	**75.7**±5.1	63.5±7.0	57.8±4.4
	EyeCLIP	69.0±5.5	60.2±2.4	50.0±0.7	50.0±0.7
	RetiZero	86.0±4.8	71.7±2.5	52.4±3.1	54.4±2.1
	VisionUnite	88.1±3.5	67.0±1.3	51.5±2.2	51.6±1.5
EDDFS_DR	RETFound-MAE	82.7±2.0	65.2±1.8	56.8±3.1	54.0±1.8
	RETFound-DINOv2	86.8±1.6	75.5±1.3	**83.4**±1.9	**70.2**±1.4
	FLAIR	**91.4**±1.8	**82.0**±1.3	76.9±4.5	66.9±2.7
	EyeCLIP	78.8±2.3	68.5±1.5	50.5±0.9	51.4±1.4
	RetiZero	87.7±1.4	77.5±1.6	52.6±1.7	57.2±1.1
	VisionUnite	83.8±2.9	68.8±1.2	56.0±2.0	57.5±1.7
Yangxi_DR	RETFound-MAE	**94.7**±1.4	63.8±1.6	63.9±8.3	55.7±1.2
	RETFound-DINOv2	83.1±1.9	58.7±1.0	**70.5**±2.2	55.9±0.6
	FLAIR	72.6±4.2	54.1±1.5	56.7±3.0	51.1±1.0
	EyeCLIP	83.5±2.2	59.2±1.0	63.6±5.5	**57.7**±1.1
	RetiZero	90.2±2.1	**64.1**±1.1	57.8±4.9	53.0±1.9
	VisionUnite	75.0±2.8	57.2±1.6	52.9±3.1	51.8±1.6

Table 10: Accuracy (%) on glaucoma tasks with frozen encoders and clean-trained linear probes. Values are means over ten folds ± the half-width of the 95% t-based confidence interval. Artifact conditions pool predictions across five artifact types at two severities; A+G combines these artifacts with grayscale. Bold marks the highest accuracy across encoders for each dataset and test condition, using unrounded values.

Dataset	Encoder	Clean	Artifact	Grayscale	A+G
AnwaraHamida_Glaucoma	RETFound-MAE	72.0±3.1	54.0±0.3	50.9±1.7	51.7±0.7
	RETFound-DINOv2	76.0±3.2	62.5±1.6	**74.7**±3.2	58.6±1.5
	FLAIR	**81.1**±3.7	**64.1**±1.7	71.8±6.0	**58.9**±1.2
	EyeCLIP	70.2±3.8	58.7±1.8	49.4±1.8	49.7±1.7
	RetiZero	78.9±2.4	60.7±1.5	58.2±6.6	52.8±2.4
	VisionUnite	77.5±3.5	62.4±1.2	56.5±3.7	52.4±1.1
EDDFS_Glaucoma	RETFound-MAE	70.6±2.7	54.0±1.2	56.5±4.1	53.0±1.4
	RETFound-DINOv2	75.5±3.0	64.7±1.5	67.0±4.3	59.1±1.7
	FLAIR	**83.9**±1.6	**70.5**±1.3	**79.1**±3.1	**68.4**±2.0
	EyeCLIP	74.8±2.3	60.2±1.0	58.9±4.8	55.5±1.4
	RetiZero	82.4±2.6	67.5±0.9	70.7±4.5	60.1±0.8
	VisionUnite	70.4±3.1	60.2±1.1	55.9±4.2	51.6±1.2
G1020	RETFound-MAE	48.8±7.1	50.5±2.2	50.6±3.6	49.8±1.5
	RETFound-DINOv2	53.8±7.9	50.6±3.6	**52.7**±6.1	**51.3**±2.9
	FLAIR	56.8±5.7	51.4±2.2	50.9±1.8	50.7±1.1
	EyeCLIP	50.6±3.8	49.0±1.9	50.3±3.4	50.5±2.3
	RetiZero	51.4±5.0	50.7±2.2	49.4±4.7	49.0±2.1
	VisionUnite	**95.3**±2.5	**58.2**±1.8	50.6±2.2	49.9±1.4
HarvardGlaucoma_1544	RETFound-MAE	84.6±4.5	72.1±3.2	56.6±4.8	52.3±1.4
	RETFound-DINOv2	87.4±5.0	**75.7**±2.3	**81.5**±4.0	**64.2**±2.9
	FLAIR	79.6±3.4	62.1±2.3	64.1±8.1	57.6±3.2
	EyeCLIP	84.3±4.6	73.4±3.3	49.7±0.9	50.1±0.9
	RetiZero	**88.0**±5.9	74.9±3.5	50.0±0.5	50.0±0.5
	VisionUnite	85.8±3.6	72.1±2.8	57.6±3.5	53.6±1.5
PAPILA	RETFound-MAE	64.3±8.3	54.3±2.8	53.5±5.3	51.3±1.4
	RETFound-DINOv2	69.7±7.2	56.7±3.9	57.1±4.7	54.4±2.1
	FLAIR	**79.7**±6.3	**63.8**±4.2	**61.3**±4.4	**57.3**±2.7
	EyeCLIP	54.0±9.5	52.3±4.2	49.7±4.4	49.9±1.8
	RetiZero	72.3±4.4	59.5±2.1	58.5±7.6	53.3±1.8
	VisionUnite	58.4±7.6	52.0±1.9	48.4±2.0	49.0±1.3

Table 11: Accuracy (%) on age-related macular degeneration tasks with frozen encoders and clean-trained linear probes. Values are means over ten folds ± the half-width of the 95% t-based confidence interval. Artifact conditions pool predictions across five artifact types at two severities; A+G combines these artifacts with grayscale. Bold marks the highest accuracy across encoders for each dataset and test condition, using unrounded values.

Dataset	Encoder	Clean	Artifact	Grayscale	A+G
ADAM400	RETFound-MAE	79.4±10.1	57.5±3.2	70.0±7.8	54.8±2.5
	RETFound-DINOv2	84.4±4.8	69.5±5.0	73.8±8.9	61.8±3.9
	FLAIR	83.1±7.9	**73.2**±5.0	**81.2**±11.2	**64.8**±5.0
	EyeCLIP	63.7±5.5	60.0±3.5	50.6±3.3	50.1±1.4
	RetiZero	85.0±6.0	73.1±5.2	69.4±9.3	52.9±2.3
	VisionUnite	**96.2**±3.1	65.4±4.8	50.0±2.1	50.2±0.6
AMDNet23_AMD	RETFound-MAE	85.9±4.0	65.6±2.5	83.1±5.1	64.3±2.4
	RETFound-DINOv2	**91.7**±2.9	75.5±1.9	**86.8**±4.0	67.7±3.8
	FLAIR	90.2±3.3	73.4±3.7	82.5±2.5	**74.4**±2.0
	EyeCLIP	85.0±3.1	71.8±1.7	71.4±6.1	62.0±2.5
	RetiZero	90.4±6.0	**80.4**±3.1	77.2±4.8	71.2±2.9
	VisionUnite	89.9±4.2	72.1±3.8	62.2±7.2	52.9±1.6
EDDFS_AMD	RETFound-MAE	62.7±2.5	55.6±2.2	50.0±0.9	50.5±1.3
	RETFound-DINOv2	70.7±3.3	62.1±2.1	59.0±3.4	56.6±2.1
	FLAIR	70.9±2.7	**64.0**±1.2	**64.8**±3.9	**58.5**±0.9
	EyeCLIP	62.4±2.5	58.5±2.0	50.0±0.0	50.0±0.1
	RetiZero	**72.1**±3.2	61.3±2.1	50.4±0.5	52.1±0.6
	VisionUnite	67.4±3.4	58.9±1.8	50.5±1.0	51.0±0.9

Table 12: Accuracy (%) on pathological myopia tasks with frozen encoders and clean-trained linear probes. Values are means over ten folds ± the half-width of the 95% t-based confidence interval. Artifact conditions pool predictions across five artifact types at two severities; A+G combines these artifacts with grayscale. Bold marks the highest accuracy across encoders for each dataset and test condition, using unrounded values.

Dataset	Encoder	Clean	Artifact	Grayscale	A+G
PALM_Myopia	RETFound-MAE	93.1±5.6	67.0±6.1	86.1±3.7	61.4±4.5
	RETFound-DINOv2	94.9±2.6	**83.1**±2.3	**93.3**±2.1	**74.4**±2.9
	FLAIR	93.6±3.9	80.3±2.8	52.1±2.6	53.1±2.0
	EyeCLIP	90.7±4.4	81.6±3.0	50.0±0.8	50.0±0.8
	RetiZero	96.5±2.2	82.2±2.5	91.4±2.9	67.8±3.0
	VisionUnite	**100.0**±0.0	82.4±1.0	45.7±4.0	44.7±2.5

Table 13: Accuracy (%) on multi-disease screening tasks with frozen encoders and clean-trained linear probes. Values are means over ten folds ± the half-width of the 95% t-based confidence interval. Artifact conditions pool predictions across five artifact types at two severities; A+G combines these artifacts with grayscale. Bold marks the highest accuracy across encoders for each dataset and test condition, using unrounded values.

Dataset	Encoder	Clean	Artifact	Grayscale	A+G
FIVES_800	RETFound-MAE	**94.5**±3.1	64.0±3.5	52.0±4.7	55.6±1.8
	RETFound-DINOv2	87.0±5.5	70.9±3.7	**74.7**±5.6	**64.2**±2.8
	FLAIR	88.4±3.1	69.7±2.6	63.3±11.7	57.5±6.3
	EyeCLIP	73.7±6.4	63.9±2.7	58.9±3.2	54.0±1.9
	RetiZero	93.8±4.1	**73.4**±3.3	69.9±5.9	59.3±2.0
	VisionUnite	85.9±4.6	65.2±2.3	65.4±6.5	53.0±2.0
JSIEC	RETFound-MAE	72.3±13.2	55.7±3.6	53.6±10.8	53.3±4.5
	RETFound-DINOv2	79.6±12.2	66.7±5.9	**73.6**±12.9	55.1±3.6
	FLAIR	84.1±7.7	**72.6**±4.3	61.6±8.6	**59.1**±6.7
	EyeCLIP	64.6±11.9	57.5±6.8	50.0±5.4	50.8±5.5
	RetiZero	85.7±7.0	67.5±3.3	64.5±9.8	53.2±4.6
	VisionUnite	**86.8**±6.4	68.4±2.7	62.5±12.5	59.0±4.9
RFMiD_Disease	RETFound-MAE	70.6±4.0	59.3±2.0	59.5±2.9	54.5±1.6
	RETFound-DINOv2	74.9±2.8	65.4±2.1	**72.7**±3.6	**62.5**±2.2
	FLAIR	**76.0**±3.6	**66.1**±1.8	69.4±5.3	61.7±2.7
	EyeCLIP	69.0±4.8	62.6±2.5	56.6±3.5	54.0±1.3
	RetiZero	75.2±2.7	59.4±2.0	59.4±3.9	55.4±1.4
	VisionUnite	75.7±3.0	62.0±1.6	51.1±2.0	50.0±1.1

F Disease-Type Breakdown

Table 14 summarizes clean-trained linear-probe performance by disease category. We average the ten folds for each task and then give each task equal weight within its category. Clean accuracy and the artifact and grayscale gaps show how performance changes across the five categories. The pathological myopia category contains only PALM, so its results represent one dataset.

RETFound-DINOv2 has the smallest grayscale gap in four categories, while FLAIR has the smallest gap for age-related macular degeneration. EyeCLIP has the smallest artifact gap in every category, but also the lowest clean accuracy. A smaller gap therefore does not necessarily indicate higher accuracy on corrupted images. The per-dataset accuracies in Appendix E provide the corresponding absolute performance for each test condition.

Table 14: Performance by disease category with frozen encoders and clean-trained linear probes. Values average ten folds per task and give tasks equal weight within each category. Clean denotes accuracy (%); Δ denotes clean minus corrupted accuracy in percentage points. The artifact condition pools predictions across five types at two severities. Bold marks the highest clean accuracy or smallest gap within each category, using unrounded values. DR denotes diabetic retinopathy; AMD denotes age-related macular degeneration.

Encoder	Clean	Δ Artifact	Δ Grayscale
DR (5 tasks)			
RETFound-MAE	85.0	21.9	22.4
RETFound-DINOv2	86.2	15.1	**6.2**
FLAIR	88.9	16.4	22.4
EyeCLIP	76.6	**13.6**	22.2
RetiZero	**89.0**	16.6	33.2
VisionUnite	84.4	17.5	29.5
Glaucoma (5 tasks)			
RETFound-MAE	68.0	11.0	14.4
RETFound-DINOv2	72.5	10.4	**5.9**
FLAIR	76.2	13.8	10.8
EyeCLIP	66.8	**8.1**	15.2
RetiZero	74.6	11.9	17.3
VisionUnite	**77.5**	16.5	23.7
AMD (3 tasks)			
RETFound-MAE	76.0	16.4	8.3
RETFound-DINOv2	82.3	13.3	9.1
FLAIR	81.4	11.2	**5.2**
EyeCLIP	70.4	**6.9**	13.0
RetiZero	82.5	10.9	16.8
VisionUnite	**84.5**	19.0	30.3
Pathological myopia (1 task)			
RETFound-MAE	93.1	26.1	7.0
RETFound-DINOv2	94.9	11.8	**1.6**
FLAIR	93.6	13.3	41.5
EyeCLIP	90.7	**9.1**	40.7
RetiZero	96.5	14.4	5.1
VisionUnite	**100.0**	17.6	54.3
Multi-disease (3 tasks)			
RETFound-MAE	79.2	19.5	24.1
RETFound-DINOv2	80.5	12.9	**6.9**
FLAIR	82.8	13.4	18.1
EyeCLIP	69.1	**7.8**	13.9
RetiZero	**84.9**	18.1	20.3
VisionUnite	82.8	17.6	23.1

G Weak and Strong Artifacts

Table 15 separates the pooled artifact condition of Table 1 into its weak and strong settings for clean-trained linear probes. We score the five weak views and the five strong views of each test image separately, using the same probes, splits, and aggregation as the pooled results. Weak artifacts alone reduce mean accuracy by 9.9 percentage points, and strong artifacts by 18.6 points. Every encoder loses at least 7.0 points under weak artifacts. Severity does not always order the damage: weak rotation reduces accuracy more than strong rotation, and weak views score lower than strong views in 77 of 510 encoder–dataset–artifact combinations. We report this breakdown for linear probing only, because the end-to-end results were recorded in pooled form.

Table 15: Accuracy gaps for weak and strong artifacts with frozen encoders and clean-trained linear probes. Values are averaged over ten folds and 17 equally weighted datasets. Clean is accuracy (%); gaps are clean minus corrupted accuracy in percentage points. Weak and strong each pool five artifact views per test image. The lower block averages the six encoders for each artifact type.

Encoder	Clean	Weak gap	Strong gap
RETFound-MAE	77.9	16.1	19.0
RETFound-DINOv2	81.0	7.0	18.6
FLAIR	83.0	9.0	19.0
EyeCLIP	72.1	7.3	11.7
RetiZero	83.3	9.2	19.5
VisionUnite	83.0	11.1	23.9
Mean	80.1	9.9	18.6
By artifact type (mean over encoders)			
Bias field		7.9	16.3
Crop		8.8	21.7
Motion		9.1	24.3
Noise		10.0	20.7
Rotation		14.0	10.0

H End-to-End Transfer Between Artifacts

Table 16 reports the full single-artifact transfer matrix for the experiment in Sec. 5.2. We fine-tune all six encoders on ADAM, APTOS, FIVES, G1020, MMRDR, and PALM. Each training method includes clean images and both severities of one artifact, with the same view composition in validation. Rows identify the training artifact and columns identify the test artifact. Diagonal entries measure performance after direct exposure; off-diagonal entries measure transfer to artifacts excluded from training and validation.

Direct exposure gives the highest mean accuracy for every test artifact. The advantage over the average of the four single-other-artifact training conditions is largest for noise (11.1 percentage points) and cropping (9.0 points), and smallest for motion (3.6 points). However, motion also has the lowest accuracy after direct exposure, at 72.5%. Its smaller transfer gap therefore accompanies lower performance even when motion is included during training. These results show that exposure to one artifact does not provide the same gains across other artifact types.

Table 16: End-to-end transfer between individual artifacts on the six-dataset subset. Training and validation use clean images and both severities of the row artifact. Entries report test accuracy (%), averaged over ten folds, six equally weighted datasets, and six encoders, with both severities of the column artifact pooled at evaluation. Bold marks the highest accuracy in each test-artifact column of the 5×5 matrix; these entries are also the diagonal direct-exposure results. The rightmost column averages transfer to the other four test artifacts. The bottom Unseen avg. row averages performance after training on each of the other four artifacts. Gap is diagonal accuracy minus this column average, in percentage points, calculated before rounding.

Trained on	Tested on					
	Bias field	Crop	Motion	Noise	Rotation	Unseen avg.
Bias field	**82.6**	69.4	67.9	69.9	76.5	70.9
Crop	76.2	**79.9**	69.7	68.8	77.2	72.9
Motion	76.6	72.5	**72.5**	68.5	79.1	74.2
Noise	76.2	70.3	68.2	**80.1**	75.3	72.5
Rotation	76.3	71.1	69.6	68.9	**83.4**	71.5
Unseen avg.	76.3	70.8	68.9	69.0	77.0	
Gap	6.3	9.1	3.6	11.1	6.4	

Table 17: Training-image comparisons with frozen encoders and linear probes. Values average ten folds for each task, then weight the 17 tasks and six encoders equally. All augmented settings include clean images. Standard and artifact augmentation each add 10 views per image; grayscale adds one. Accuracy is in percent; Δ_A and Δ_G are clean minus artifact and grayscale accuracy, respectively, in percentage points. Bold marks the highest accuracy or smallest gap in each column.

Training images	Accuracy (%)			Gap (points)	
	Clean	Artifacts	Grayscale	Δ_A	Δ_G
Clean only	80.1	65.8	62.1	14.3	18.0
Clean + standard aug.	80.4	69.1	63.8	11.3	16.6
Clean + artifacts	**82.1**	**75.4**	63.6	**6.7**	18.5
Clean + grayscale	80.7	66.1	**77.3**	14.6	**3.4**

I Artifact and Grayscale Training with Frozen Encoders

We repeat the augmentation comparisons with all six encoders frozen across all 17 tasks, complementing the end-to-end experiments in Secs. 5.1 and 5.3. Table 17 summarizes clean, artifact, and grayscale accuracy and the corresponding robustness gaps for the four training settings. Each linear probe is trained on clean features alone or on clean features combined with standard-augmentation, artifact, or grayscale features. Standard and artifact augmentation each add 10 views per image; grayscale training adds one. We use the data partitions and probe optimization settings from Sec. 3.4. Feature-standardization statistics are fit on the training features for each method, and validation uses the same view composition as training. Test images and their views remain excluded from training and validation.

The standard-augmentation control uses flips, rotation, resized crops, color jitter, and blur. It improves mean artifact accuracy from 65.8% to 69.1% and reduces the artifact gap from 14.3 to 11.3 percentage points. Mean clean accuracy increases from 80.1% to 80.4%. As in the end-to-end comparison, standard augmentation provides some benefit, but substantial artifact gaps remain.

Adding artifact-corrupted images produces larger gains. Mean artifact accuracy reaches 75.4%, while clean accuracy rises to 82.1%. The mean artifact gap falls to 6.7 points, a reduction of 7.6 points from clean-only training and 4.6 points from standard augmentation. These gains show that the frozen representations support better classification of corrupted images when the probe is trained with artifact views. The comparison includes the corresponding changes in feature standardization and validation. Both augmented methods use the same number of views per epoch, although early stopping can produce different total numbers of training steps.

The gains do not extend equally between artifacts and grayscale. Artifact training leaves a mean grayscale gap of 18.5 points, compared with 18.0 points for clean-only training. Grayscale training reduces this gap to 3.4 points, but leaves an artifact gap of 14.6 points. Its mean clean accuracy is 80.7%, compared with 80.1% for the clean-only baseline. Thus, both interventions improve mean clean accuracy, while their robustness gains primarily concern the image changes used during training. As with end-to-end fine-tuning, these results support evaluating artifact robustness and color sensitivity separately.

J Checks on Two Design Choices

We examine two settings used in the benchmark: feature standardization for linear probing and batch size for end-to-end fine-tuning. Table 18 compares all six frozen encoders with and without standardization. Table 19 reports the changes associated with using a shared fine-tuning batch size of 32.

For the standardization check, we train linear probes on MMRDR, G1020, ADAM, PALM, and JSIEC, with one task from each disease category. Standardization uses the mean and standard deviation of each feature dimension in the training partition. We retain the same data splits and probe training settings in both conditions.

Table 18: Effect of feature standardization on frozen-probe clean accuracy (%), averaged over five datasets, one per disease category, for six encoders. On and Off indicate whether features are standardized. Change is Off minus On, in percentage points; negative values indicate lower accuracy without standardization. Differences use unrounded results.

Encoder	Clean (on)	Clean (off)	Change
RETFound-MAE	74.2	71.3	-2.9
RETFound-DINOv2	79.1	79.0	-0.1
FLAIR	82.1	79.4	-2.8
EyeCLIP	67.9	53.3	-14.6
RetiZero	80.8	76.1	-4.7
VisionUnite	91.6	83.4	-8.2

Table 19: Fine-tuning batch-size settings and reported changes in performance. Arrows show the original and shared batch sizes. Changes are shared minus original values, in percentage points, calculated before rounding. Clean denotes clean accuracy; grayscale gap denotes clean minus grayscale accuracy. RETFound-DINOv2 uses batch size 32 throughout; its zero entries indicate an unchanged setting, not a separate rerun.

Encoder	Batch	Change in clean	Change in grayscale gap
RETFound-MAE	$32 \rightarrow 32$	$+0.0$	$+0.0$
RETFound-DINOv2	$32 \rightarrow 32$	$+0.0$	$+0.0$
FLAIR	$24 \rightarrow 32$	-0.1	$+0.1$
EyeCLIP	$64 \rightarrow 32$	$+0.6$	$+0.5$
RetiZero	$32 \rightarrow 32$	$+0.0$	$+0.0$
VisionUnite	$8 \rightarrow 32$	-1.8	-0.8

The effect differs substantially across encoders (Table 18). RETFound-DINOv2 has nearly identical clean accuracy with and without standardization, at 79.1% and 79.0%, respectively. The other five encoders lose 2.8–14.6 percentage points without standardization. EyeCLIP shows the largest decrease, from 67.9% to 53.3%. Thus, feature standardization provides little benefit for RETFound-DINOv2 in this comparison, but substantially affects several other encoders. We use the same standardization procedure across encoders to keep the linear-probe protocol consistent.

For fine-tuning, the shared batch size replaces values of 24 for FLAIR, 64 for EyeCLIP, and 8 for VisionUnite (Table 19). The reported changes in clean accuracy are -0.1, $+0.6$, and -1.8 percentage points, respectively. RETFound-MAE and RetiZero already use batch size 32. RETFound-DINOv2 also uses 32 throughout, so its row records an unchanged setting rather than a new batch-size experiment. All main fine-tuning comparisons use batch size 32 to keep this setting consistent across encoders.

K Reproducibility

We will release the benchmark code, dataset manifests, split indices, and saved per-fold results with the camera-ready version. The release will explain how to inspect the results, generate the summary tables, and run the experiments.

All encoders use shared data partitions and generated image views before model-specific preprocessing. Corruption seeds are determined by a fixed base seed, the image identifier, and the test condition. Appendix C describes the grouping, training, validation, and aggregation procedures.

The main comparison (Table 1) and primary end-to-end interventions (Table 4) each cover 102 encoder–dataset pairs. The augmentation control and transfer analyses (Tables 2, 3, and 16) cover all six encoders on six datasets. Each completed evaluation condition contains ten folds.

We will also share the benchmark dataset with the camera-ready version, subject to the source datasets' licenses and access conditions. For sources that cannot be redistributed, we will provide acquisition and preparation instructions.

AI USE STATEMENT

We used AI tools, including Claude and ChatGPT, to assist with code debugging and refinement, manuscript revision, and figure generation. We did not use AI agents or large language models (LLMs) for research ideation or problem solving. We did not rely solely on LLM-generated revisions; we reviewed and revised the content to ensure accuracy and consistency. We take full responsibility for the content of this manuscript.