LEARNING FROM HETEROGENEOUS ANNOTATION MECHANISMS FOR UNIFIED RETINAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning a unified retinal predictor from datasets collected for different clinical purposes is complicated by a fundamental ambiguity: an unrecorded label is not necessarily clinically negative. Because annotations follow heterogeneous, source-specific protocols, flattening such datasets into a common binary label matrix can introduce systematic false-negative supervision. Masking unobserved entries prevents this error, but uses label observability only to gate the loss and leaves both the structured observation process and uneven supervision across source–label blocks unmodeled. We introduce Annotation-Mechanism-Aware learning (AMA), which separates clinical prediction from label observability. AMA combines an auxiliary label-observability objective with source–label schema balancing, while requiring neither source identity nor observation masks at inference. We evaluate AMA on 252,116 fundus photographs from 54 public datasets whose annotations are harmonized into a 98-label clinical vocabulary. With the same pre-trained retinal encoder and training budget, AMA attains a macro-AUPRC of 0.802 (95% CI: 0.788–0.823), compared with 0.545 (95% CI: 0.538–0.572) for the matched masked-supervision Baseline. These gains persist across four retinal backbones spanning masked image modeling, self-distillation, and vision–language pre-training, with improved per-label AUPRC estimates for 96 labels. The results support treating label observability as part of the learning problem when pooling medical datasets, rather than assuming that the pooled label matrix is complete.

1 INTRODUCTION

Public retinal-image datasets constitute a valuable source of clinical supervision. Collectively, they span diverse populations, acquisition devices, diseases, severity scales, and observable findings, and have enabled the development of increasingly capable retinal foundation models (Zhou et al., 2023; Qiu et al., 2024; Silva-Rodríguez et al., 2024; Du et al., 2024). Pooling these datasets appears to offer a direct route toward a unified predictor that can recognize many retinal conditions from a single image. The difficulty, however, is not merely that the constituent datasets use different label names. They were created for different clinical purposes and therefore record different parts of the underlying clinical state.

A label that is absent from a medical dataset is not necessarily negative. For example, a diabetic-retinopathy dataset may systematically record disease severity while leaving glaucoma, age-related macular degeneration, or individual retinal findings unassessed (Figure 1a). Task-specific negatives also require care: AIROGS defines "no referable glaucoma" irrespective of comorbidities such as diabetic retinopathy or age-related macular degeneration (de Vente et al., 2024). Thus, the "Normal" example in Figure 1a does not establish comprehensive ocular health. Once such datasets are flattened into a common binary label matrix, conventional training can silently turn these unobserved entries into negative targets. This can introduce systematic false-negative supervision tied to source annotation protocols (Figure 1d). We refer to this phenomenon as *annotation-mechanism bias*: the learned predictor can reflect which labels a source chose to assess instead of the latent clinical state that the labels are intended to describe.

Recent retinal foundation models improve visual representations through masked image modeling, self-distillation, and vision–language pre-training (Zhou et al., 2023; Yu et al., 2024; Shi et al., 2025; Wang et al., 2025; Wu et al., 2024). Related work further leverages multiple textual labels and clinical

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.

descriptions across semantic granularities (Li et al., 2026). These advances enrich observed visual or semantic supervision but do not explicitly represent label observability under source-specific protocols. Our setting addresses a different aspect of supervision: whether an apparent label absence represents a confirmed negative or an unobserved state under a source-specific protocol.

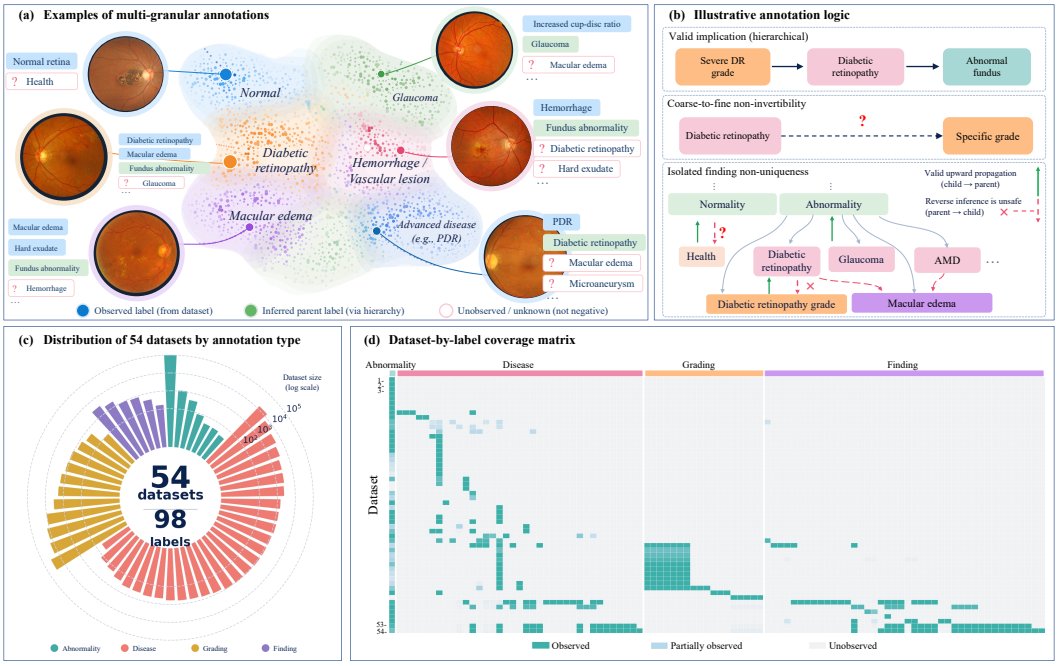

Figure 1: Heterogeneous annotations and partial observability in multi-source retinal data. **(a)** Examples of heterogeneous label granularity and multiplicity; a task-specific "Normal" label need not imply comprehensive ocular health. Observed, hierarchy-inferred, and unobserved labels are distinguished. **(b)** Valid upward mappings do not justify coarse-to-fine or finding-to-disease reverse inference. **(c)** Scale and annotation types of 54 datasets (252,116 images; 98 labels). **(d)** Source–label coverage reveals structured partial observability: unobserved labels are not negatives.

Related multi-label objectives address class imbalance, missing labels, and single-positive supervision (Ridnik et al., 2021; Wu et al., 2020; Kim et al., 2022; Zhou et al., 2022; Cole et al., 2021; Ben-Baruch et al., 2022). Yet common formulations treat missingness as an unobserved entry pattern to be masked, corrected, or estimated. In a multi-source medical corpus, missingness is instead highly structured: the probability that label c is observed depends on the source, its clinical objective, and its annotation schema. Merely masking unobserved entries prevents one error—interpreting each as a negative—but uses the observation mask only to gate the clinical loss and leaves its source-dependent structure unmodeled. Densely annotated source–label blocks can therefore dominate the remaining supervision.

We introduce *Annotation-Mechanism-Aware learning* (AMA), a framework for learning a unified retinal predictor from source-structured, partially observed multi-label supervision. AMA explicitly separates latent clinical status from label observability. It uses the observation mask as an auxiliary learning signal and balances observed clinical supervision across source–label annotation blocks, while never imputing an unobserved label as positive or negative. At inference, only the shared image encoder and clinical prediction head are retained; source identity and observation masks are not required.

We evaluate AMA after harmonizing 54 public retinal datasets comprising 252,116 fundus photographs into a 98-label clinical vocabulary. Under the same pre-trained retinal encoder and training budget, AMA achieves a macro-AUPRC of 0.802 (95% CI: 0.788–0.823), compared with 0.545 (0.538–0.572) for the matched masked-supervision Baseline. Improvements also persist across four retinal backbones representing masked image modeling, self-distillation, and vision–language pre-training, and extend across disease, grading, and finding labels. Our contributions are:

- We formulate multi-source retinal learning as prediction under heterogeneous annotation mechanisms, explicitly distinguishing unobserved labels from confirmed negatives.
- We propose AMA, which combines observation-aware representation learning with source–label schema-balanced clinical supervision, while requiring neither source identity nor observation masks at inference.
- We evaluate AMA across 54 datasets, 98 labels, multiple competing objectives, component ablations, and four retinal foundation-model backbones, together with label-level analyses by annotation category and training support.

2 Problem Formulation

2.1 Multi-source learning with partially observed labels

Consider a collection of retinal image datasets acquired under different clinical tasks and annotation protocols. We map their source-specific vocabularies to a common set of C clinical labels $\mathcal{C} = \{1, \dots, C\}$. The resulting training set is $\mathcal{D} = \{(x_i, s_i, y_i)\}_{i=1}^{N}$, where x_i is an image, $s_i \in \{1, \dots, S\}$ identifies its data source, and $y_i \in \{0, 1, \mathtt{U}\}^C$ is a partially observed multi-label target. Here, 1 and 0 denote an observed positive and an observed negative, respectively, whereas U denotes an unobserved label. U is not a clinical negative: it states that the corresponding clinical state was not assessed under the available annotation protocol.

Following the distinction between data and their observation process (Rubin, 1976), let $z_{ic} \in \{0, 1\}$ denote the latent clinical state of image i for label c, and let $o_{ic} \in \{0, 1\}$ indicate whether that state is observed. The available target is generated as

$$y_{ic} = \begin{cases} z_{ic}, & o_{ic} = 1, \\ \mathtt{U}, & o_{ic} = 0. \end{cases} \tag{1}$$

In the pooled corpus considered here, observation is source-structured rather than a uniform random deletion process. A dataset collected for diabetic retinopathy grading, for example, can systematically observe retinopathy and its severity while leaving unrelated diseases or retinal findings unassessed. We therefore allow the observation distribution to depend on both source and label,

$$q_{sc} := \Pr(o_{ic} = 1 \mid s_i = s), \qquad c \in \mathcal{C}, \tag{2}$$

without assuming that the complete latent target vector z_i is available. The matrix $Q = [q_{sc}] \in [0, 1]^{S \times C}$ is a conceptual description of the heterogeneous source–label annotation schema. At the block level, $q_{sc} = 0$, $0 < q_{sc} < 1$, and $q_{sc} = 1$ correspond to unobserved, partially observed, and fully observed blocks, respectively, as visualized in Figure 1d. Operationally, AMA uses observation masks and empirical source–label support (Section 3) rather than estimating Q explicitly.

2.2 Why masking unobserved labels is insufficient

A natural baseline is to exclude unobserved entries from the clinical classification loss. Let $f_\theta(g_\phi(x_i))_c$ be the logit for label c. Masked binary cross-entropy minimizes

$$\mathcal{L}_{\text{mask}}(\phi, \theta) = \frac{\sum_{i=1}^{N} \sum_{c=1}^{C} o_{ic}[-\rho_c y_{ic} \log p_{ic} - (1 - y_{ic}) \log(1 - p_{ic})]}{\sum_{i=1}^{N} \sum_{c=1}^{C} o_{ic}}. \tag{3}$$

Here $p_{ic} = \sigma(f_\theta(g_\phi(x_i))_c)$, and $\rho_c = (N_c^- + 1)/(N_c^+ + 1)$ compensates for within-label imbalance, where N_c^+ and N_c^- are the numbers of observed positive and negative training targets for label c, respectively. Masked terms are omitted without evaluating U numerically. However, this objective uses label observability only as a gate and otherwise regards all observed entries as exchangeable. Let $n_{sc} = \sum_{i=1}^{N} \mathbf{1}[s_i = s] o_{ic}$ be the number of observed training targets for source–label pair (s, c). For a fixed label c, the relative contribution of source s to Equation 3 is driven by n_{sc}. Consequently, source–label blocks with abundant annotations can dominate the gradient, while sparsely observed blocks contribute less. Masking also leaves the structured observation pattern unmodeled, although that pattern identifies which supervision was available for each image.

Our goal is therefore not to infer that every unobserved entry is positive or negative. Rather, we seek a shared predictor over the unified label space that (i) preserves the distinction between unobserved and observed-negative targets, (ii) explicitly represents label observability, and (iii) mitigates the disproportionate influence of densely annotated source–label blocks. Masked clinical prediction enforces (i), observation-aware learning addresses (ii), and schema-balanced supervision addresses (iii), as detailed in Section 3.

3 METHOD

3.1 OVERVIEW

AMA learns a unified clinical predictor while accounting for how supervision is distributed across annotation protocols, as illustrated in Figure 2. Given image x_i, a shared retinal encoder g_ϕ produces a representation $h_i = g_\phi(x_i)$. Two prediction heads operate on this representation: a clinical head f_θ predicts the C retinal labels, and an observation head r_ψ predicts which label positions are observed. In parallel, fixed coefficients derived from source–label observation support balance the contributions of observed source–label blocks to the clinical objective. AMA is trained with

$$\mathcal{L}_{\text{AMA}} = \mathcal{L}_{\text{clinical}} + \lambda_{\text{obs}} \mathcal{L}_{\text{obs}}, \tag{4}$$

where λ_{obs} controls the contribution of the label-observability objective. The two terms share the encoder but supervise different aspects of the training signal: clinical status and label observability.

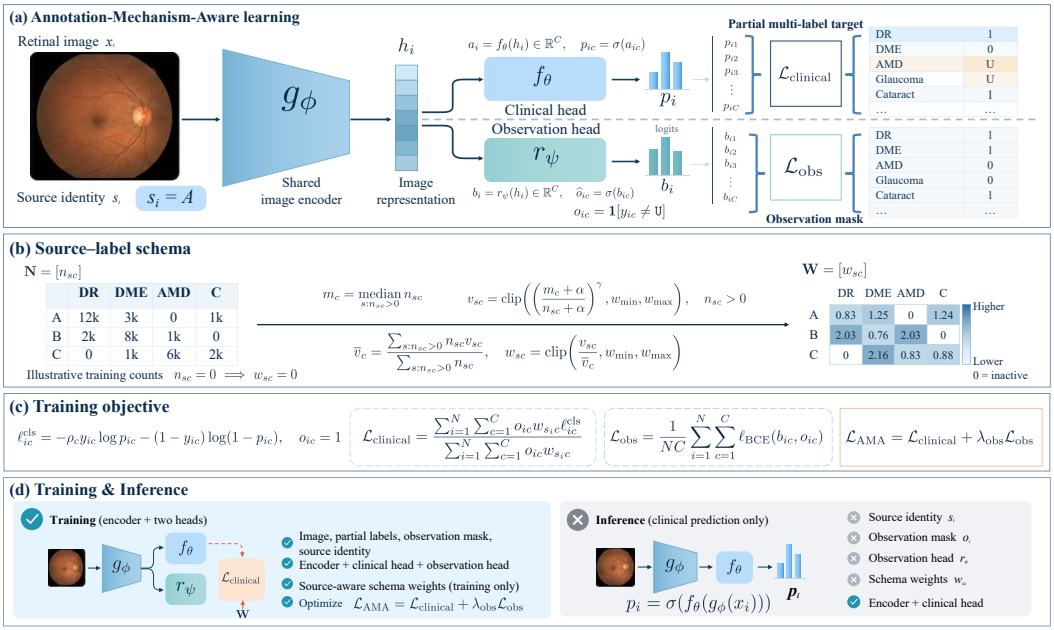

Figure 2: Overview of AMA. **(a)** A shared encoder feeds clinical and observation heads to predict clinical labels and their observability, respectively. **(b)** Training-set source–label support defines fixed coefficients for schema-balanced supervision. **(c)** Joint optimization combines masked, schema-balanced clinical prediction with observation prediction, without treating unobserved labels as negatives. **(d)** Both heads train the encoder; inference retains only the encoder and clinical head, requiring neither source identity nor observation masks.

3.2 MASKED CLINICAL PREDICTION

The clinical head produces logits

$$a_i = f_\theta(h_i) \in \mathbb{R}^C, \qquad p_{ic} = \sigma(a_{ic}). \tag{5}$$

For an observed target, we use positive-class-weighted binary cross-entropy

$$\ell_{ic}^{\text{cls}} = -\rho_c y_{ic} \log p_{ic} - (1 - y_{ic}) \log(1 - p_{ic}), \qquad o_{ic} = 1, \tag{6}$$

where ρ_c compensates for the positive–negative imbalance within label c. Specifically, $\rho_c = (N_c^- + 1)/(N_c^+ + 1)$, using the observed positive and negative counts in the training split. Unobserved entries are excluded through o_{ic} and are never converted into negative clinical targets. Setting $w_{sc} = 1$ and $\lambda_{\text{obs}} = 0$ recovers the masked clinical objective in Equation 3.

3.3 Observation-aware representation learning

Masking prevents incorrect negative supervision but uses the observation mask only to gate the clinical loss. AMA instead makes label observability an auxiliary prediction task. The observation head produces

$$b_i = r_\psi(h_i) \in \mathbb{R}^C, \qquad \widehat{o}_{ic} = \sigma(b_{ic}), \tag{7}$$

and is supervised at every label position by

$$\mathcal{L}_{\text{obs}} = \frac{1}{NC} \sum_{i=1}^{N} \sum_{c=1}^{C} \ell_{\text{BCE}}(b_{ic}, o_{ic}). \tag{8}$$

Here $\ell_{\text{BCE}}(b, o) = -o \log \sigma(b) - (1 - o) \log(1 - \sigma(b))$. Because both heads share g_ϕ, this objective provides auxiliary supervision about which clinical targets were observed under the annotation process. The target o_{ic} describes whether a clinical state was recorded, not whether that state is positive. Thus, $\mathcal{L}_{\text{obs}}$ neither imputes the latent target z_{ic} nor turns an unobserved label into clinical supervision.

3.4 Source–label schema-balanced supervision

Even after unobserved entries are masked, observed supervision can be highly uneven across source–label blocks. We construct a fixed weight matrix $\mathbf{W} = [w_{sc}] \in \mathbb{R}_{\geq 0}^{S \times C}$ from the training split only. Recall that n_{sc} is the observed support for source s and label c. For each label, we first compute the median support among active sources,

$$m_c = \underset{s:n_{sc}>0}{\text{median}}\, n_{sc}. \tag{9}$$

For active source–label blocks, an initial support-balancing weight is

$$v_{sc} = \text{clip}\left(\left(\frac{m_c + \alpha}{n_{sc} + \alpha} \right)^\gamma, w_{\min}, w_{\max} \right), \qquad n_{sc} > 0, \tag{10}$$

where α stabilizes small counts and γ controls the strength of balancing. We then normalize the weights separately for each label using the observed support,

$$\overline{v}_c = \frac{\sum_{s:n_{sc}>0} n_{sc} v_{sc}}{\sum_{s:n_{sc}>0} n_{sc}}, \qquad w_{sc} = \text{clip}\left(\frac{v_{sc}}{\overline{v}_c}, w_{\min}, w_{\max} \right). \tag{11}$$

Inactive blocks with $n_{sc} = 0$ receive $w_{sc} = 0$. The schema-balanced clinical objective is

$$\mathcal{L}_{\text{clinical}} = \frac{\sum_{i=1}^{N} \sum_{c=1}^{C} o_{ic} w_{s_i c} \ell_{ic}^{\text{cls}}}{\sum_{i=1}^{N} \sum_{c=1}^{C} o_{ic} w_{s_i c}}. \tag{12}$$

The two weighting factors serve different purposes: ρ_c addresses positive–negative imbalance within a label, whereas w_{sc} addresses supervision imbalance across source–label blocks. Before the final clipping step, support-weighted normalization sets the observation-count-weighted mean coefficient within each label to one; clipping then bounds the influence of extreme source–label coefficients. Unlike prediction-dependent reweighting, these coefficients are computed once from training-set observation support and remain fixed during optimization. Unless varied in the sensitivity analysis, we use $\lambda_{\text{obs}} = 0.5$, $\gamma = 0.5$, $\alpha = 1$, and $(w_{\min}, w_{\max}) = (0.25, 4)$.

3.5 TRAINING AND INFERENCE

During training, gradients from both terms in Equation 4 update the shared encoder. The clinical head is optimized only from observed clinical targets using Equation 12, while the observation head is optimized from the complete binary observation mask using Equation 8. All support statistics and schema weights are computed from the training partition and then held fixed.

At inference, AMA retains the shared encoder and clinical head to produce $p_i = \sigma(f_\theta(g_\phi(x_i)))$. The observation head, source identity, observation mask, and schema weights are not required to generate clinical predictions. Consequently, observability modeling acts as a training signal rather than a source-dependent post-processing rule.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Data Preparation. We consolidate 54 publicly available retinal imaging datasets, comprising 252,116 fundus photographs collected under heterogeneous clinical tasks and annotation protocols. Source-specific annotations are harmonized into a unified vocabulary of 98 retinal labels spanning abnormality, disease, grading, and clinical findings. The harmonization preserves the distinction between confirmed negatives and labels that were not assessed by the corresponding source protocol. We use a fixed split of 176,481 images for training, 25,212 for validation, and 50,423 for testing. Label mapping is specified independently of model predictions, and all observability statistics used during training are computed from the training partition. Dataset-level statistics and the complete label dictionary are provided in the appendix.

Compared Methods. We evaluate AMA against two complementary groups of methods. The first comprises retinal foundation models learned with representative large-scale pre-training paradigms: RETFound (Zhou et al., 2023), VisionUnite (Li et al., 2025), VisionFM (Qiu et al., 2024), Ur-Found (Yu et al., 2024), RET-CLIP (Du et al., 2024), MIRAGE (Morano et al., 2025), PRETI (Lee et al., 2025), RetFiner (Fecso et al., 2025), ViLReF (Yang et al., 2024), FLAIR (Silva-Rodríguez et al., 2024), EyeCLIP (Shi et al., 2025), RetiZero (Wang et al., 2025), and KeepFIT (Wu et al., 2024). The second comprises established multi-label learning approaches designed for class imbalance, missing labels, or noisy supervision, including ASL (Ridnik et al., 2021), DB (Wu et al., 2020), LL-R and LL-Ct (Kim et al., 2022), EM (Zhou et al., 2022), ROLE (Cole et al., 2021), and CSL (Ben-Baruch et al., 2022). We additionally include a backbone-matched *Baseline* that uses the same pre-trained retinal encoder, data split, prediction head, optimization schedule, and training budget as AMA but omits observation-aware learning and schema-balanced supervision.

Training Protocol. Our primary experiments use a retinal encoder pre-trained with the DINOv2 objective (Oquab et al., 2024) and optimize all models for 100 epochs with AdamW (Loshchilov & Hutter, 2019). Model selection is performed on the validation set using macro-AUPRC, after which the selected checkpoint is evaluated once on the held-out test set. Unless an experiment explicitly varies the encoder, AMA and its matched Baseline share the same data processing, optimization schedule, and training budget.

Evaluation Protocol. We take macro-AUPRC as the primary performance measure because precision–recall evaluation is informative under substantial class imbalance (Saito & Rehmsmeier, 2015). Precision, specificity, F1 score, balanced accuracy, and Cohen's kappa provide complementary assessments of predictive performance. We report point estimates with 95% confidence intervals from 500 nonparametric bootstrap resamples (Efron & Tibshirani, 1993). Paired comparisons against the matched Baseline use two-sided normal-approximation tests based on aligned bootstrap differences. Holm adjustment is applied jointly to all paired tests across the six reported measures within each analysis table (Holm, 1979). The intervals quantify held-out-sample uncertainty. All operating thresholds are selected on the validation set and then fixed for test evaluation.

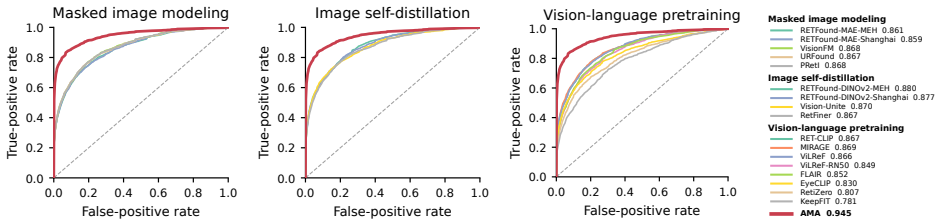

Figure 3: Macro-average ROC curves for retinal foundation models grouped by pre-training paradigm. AMA is repeated in each panel as a common reference. The rightmost panel reports the shared legend and macro-AUROC values.

Table 1: Comparison with retinal foundation models. Values are estimates with 95% bootstrap confidence intervals. Best and second-best results are shown in bold and underlined, respectively.

Method	Precision^{Mi} (95% CI)	Specificity^{Mi} (95% CI)	F1^{Ma} (95% CI)	Bal. Acc.^{Mi} (95% CI)	κ^{Ma} (95% CI)	AUPRC^{Ma} (95% CI)
RETFound-MAE-MEH	0.675 (0.672–0.679)	0.924 (0.923–0.925)	0.495 (0.475–0.511)	0.854 (0.852–0.856)	0.419 (0.398–0.436)	0.529 (0.520–0.555)
RETFound-MAE-Shanghai	0.676 (0.673–0.680)	0.924 (0.923–0.925)	0.473 (0.453–0.487)	0.856 (0.854–0.858)	0.394 (0.372–0.411)	0.526 (0.517–0.553)
RETFound-DINOv2-MEH	0.704 (0.700–0.708)	0.930 (0.929–0.931)	0.495 (0.478–0.508)	0.874 (0.872–0.876)	0.423 (0.404–0.437)	0.545 (0.538–0.572)
RETFound-DINOv2-Shanghai	0.724 (0.720–0.727)	0.938 (0.937–0.939)	0.509 (0.491–0.522)	0.868 (0.866–0.870)	0.437 (0.417–0.452)	0.547 (0.540–0.572)
VisionUnite	0.707 (0.703–0.711)	0.935 (0.934–0.936)	0.504 (0.486–0.516)	0.857 (0.855–0.859)	0.427 (0.409–0.442)	0.541 (0.534–0.570)
VisionFM	0.683 (0.679–0.686)	0.925 (0.924–0.926)	0.480 (0.460–0.492)	0.860 (0.858–0.862)	0.406 (0.387–0.421)	0.527 (0.519–0.555)
UrFound	0.679 (0.676–0.683)	0.924 (0.923–0.925)	0.481 (0.462–0.492)	0.859 (0.857–0.861)	0.406 (0.387–0.420)	0.524 (0.516–0.551)
RET-CLIP	0.696 (0.693–0.700)	0.931 (0.930–0.932)	0.489 (0.471–0.502)	0.858 (0.856–0.860)	0.412 (0.393–0.427)	0.530 (0.521–0.556)
MIRAGE	0.682 (0.679–0.686)	0.925 (0.924–0.926)	0.476 (0.458–0.488)	0.859 (0.858–0.861)	0.402 (0.383–0.416)	0.525 (0.518–0.553)
PRETI	0.678 (0.674–0.682)	0.923 (0.922–0.924)	0.482 (0.463–0.495)	0.861 (0.859–0.863)	0.409 (0.390–0.424)	0.525 (0.517–0.552)
RetFiner	0.670 (0.667–0.674)	0.920 (0.919–0.921)	0.477 (0.457–0.489)	0.862 (0.860–0.864)	0.405 (0.386–0.419)	0.524 (0.515–0.551)
ViLReF	0.694 (0.690–0.697)	0.930 (0.928–0.931)	0.492 (0.471–0.506)	0.859 (0.857–0.861)	0.413 (0.391–0.428)	0.528 (0.519–0.555)
ViLReF-RN50	0.682 (0.678–0.686)	0.925 (0.924–0.927)	0.450 (0.431–0.464)	0.858 (0.856–0.860)	0.372 (0.353–0.388)	0.487 (0.480–0.512)
FLAIR	0.677 (0.673–0.680)	0.924 (0.922–0.925)	0.447 (0.432–0.458)	0.857 (0.856–0.860)	0.367 (0.351–0.379)	0.480 (0.474–0.507)
EyeCLIP	0.716 (0.712–0.720)	0.939 (0.938–0.940)	0.463 (0.444–0.475)	0.848 (0.846–0.850)	0.383 (0.364–0.398)	0.495 (0.487–0.521)
RetiZero	0.720 (0.716–0.724)	0.940 (0.939–0.941)	0.447 (0.429–0.459)	0.851 (0.849–0.853)	0.366 (0.346–0.380)	0.446 (0.441–0.471)
KeepFIT	0.640 (0.636–0.644)	0.911 (0.909–0.912)	0.383 (0.370–0.394)	0.848 (0.846–0.850)	0.284 (0.270–0.297)	0.404 (0.399–0.426)
RETFound-DINOv2-MEH + AMA	**0.892 (0.889–0.895)**	**0.978 (0.978–0.979)**	**0.755 (0.730–0.769)**	**0.935 (0.934–0.937)**	**0.718 (0.693–0.733)**	**0.802 (0.788–0.823)**

^{Mi} Micro aggregation; ^{Ma} macro aggregation.

4.2 Comparative Experiments

Comparison with Retinal Foundation Models. We first compare existing retinal foundation models on the unified 98-label task under a common downstream protocol. Each comparator retains its pre-trained encoder, whereas AMA applies the proposed objective to RETFound-DINOv2-MEH. The comparison covers masked image modeling, image self-distillation, and vision–language pre-training. As shown in Figure 3, the strongest foundation-model comparator reaches a macro-AUROC of 0.880. AMA attains 0.945, higher than each comparator across the three pre-training groups. The same pattern holds for the complementary measures in Table 1: AMA attains a macro-AUPRC of 0.802 (95% CI: 0.788–0.823), compared with 0.547 (0.540–0.572) for the strongest foundation-model comparator, and achieves the highest point estimate on all six reported measures. Under this common downstream protocol, the comparison motivates treating supervision structure explicitly in addition to selecting a pre-trained representation.

Comparison with Multi-Label Learning Methods. We next compare AMA with methods targeting class imbalance, incomplete annotations, or noisy supervision using the same retinal encoder and optimization schedule. Figure 4a shows that AMA attains the highest macro-AUPRC point estimate among the evaluated methods. Its estimate of 0.802 (0.788–0.823) exceeds the strongest competing method, LL-R, at 0.780 (0.768–0.803), with a paired bootstrap mean improvement over the Baseline of 0.252 (0.233–0.274). The paired analysis in Figure 4b further shows positive changes relative to the Baseline across precision, specificity, F1, balanced accuracy, Cohen's kappa, and AUPRC. Table 2 reports the corresponding absolute estimates and confidence intervals. AMA obtains the highest point estimate for every reported measure under the evaluated protocol.

4.3 Ablation Studies

Component Ablation. We compare variants containing observation-aware representation learning, source–label schema balancing, or both under the same encoder and optimization protocol. Figure S1 shows that either component improves over the Baseline across the four displayed measures; all paired

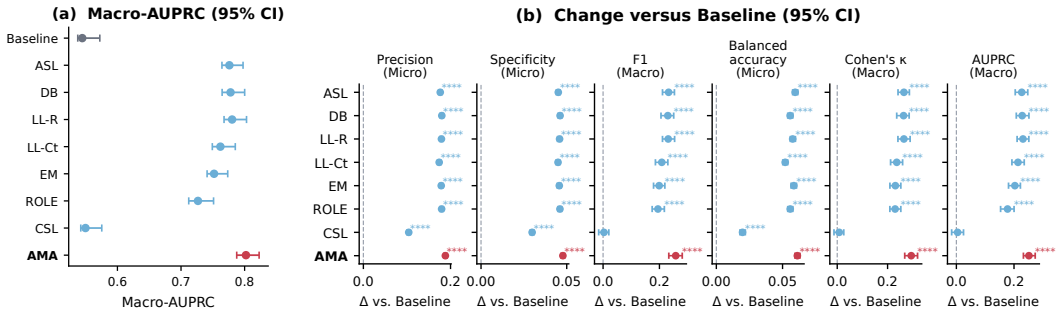

Figure 4: Comparison with multi-label learning methods. **(a)** Macro-AUPRC estimates with 95% bootstrap confidence intervals. **(b)** Paired bootstrap mean changes relative to the Baseline across six evaluation measures. Significance markers use Holm-adjusted P values.

Table 2: Comparison with multi-label learning methods under the matched DINOv2-pre-trained retinal encoder. Values are estimates with 95% bootstrap confidence intervals. Statistical comparisons use the Baseline as the reference.

Method	Precision^{Mi} (95% CI)	P_{adj}	Specificity^{Mi} (95% CI)	P_{adj}	F1^{Ma} (95% CI)	P_{adj}	Bal. Acc.^{Mi} (95% CI)	P_{adj}	κ^{Ma} (95% CI)	P_{adj}	AUPRC^{Ma} (95% CI)	P_{adj}
Baseline	0.704 (0.700–0.708)	Ref.	0.930 (0.929–0.931)	Ref.	0.495 (0.478–0.508)	Ref.	0.874 (0.872–0.876)	Ref.	0.423 (0.404–0.437)	Ref.	0.545 (0.538–0.572)	Ref.
ASL	0.880 (0.877–0.883)	****	0.975 (0.975–0.976)	****	0.730 (0.706–0.744)	****	0.934 (0.932–0.935)	****	0.689 (0.665–0.704)	****	0.776 (0.764–0.798)	****
DB	0.884 (0.881–0.887)	****	0.977 (0.976–0.977)	****	0.726 (0.705–0.741)	****	0.930 (0.928–0.931)	****	0.687 (0.665–0.703)	****	0.778 (0.765–0.800)	****
LL-R	0.883 (0.880–0.886)	****	0.976 (0.976–0.977)	****	0.728 (0.707–0.742)	****	0.932 (0.930–0.933)	****	0.689 (0.665–0.704)	****	0.780 (0.768–0.803)	****
LL-Ct	0.878 (0.875–0.881)	****	0.975 (0.975–0.976)	****	0.705 (0.682–0.718)	****	0.926 (0.924–0.928)	****	0.660 (0.637–0.674)	****	0.762 (0.749–0.785)	****
EM	0.883 (0.880–0.886)	****	0.976 (0.975–0.977)	****	0.696 (0.672–0.709)	****	0.933 (0.931–0.934)	****	0.655 (0.630–0.669)	****	0.752 (0.741–0.773)	****
ROLE	0.883 (0.881–0.886)	****	0.976 (0.976–0.977)	****	0.691 (0.669–0.706)	****	0.930 (0.928–0.931)	****	0.654 (0.631–0.670)	****	0.727 (0.712–0.751)	****
CSL	0.808 (0.804–0.811)	****	0.960 (0.959–0.961)	****	0.497 (0.482–0.509)	–	0.894 (0.892–0.896)	****	0.430 (0.413–0.443)	–	0.550 (0.542–0.575)	–
AMA	**0.892 (0.889–0.895)**	****	**0.978 (0.978–0.979)**	****	**0.755 (0.730–0.769)**	****	**0.935 (0.934–0.937)**	****	**0.718 (0.693–0.733)**	****	**0.802 (0.788–0.823)**	****

^{Mi} Micro aggregation; ^{Ma} macro aggregation. Holm-adjusted significance: ***$P < 0.001$ and ****$P < 0.0001$; "–" does not meet either threshold. Bold and underline indicate the best and second-best estimates, respectively.

comparisons remain significant after Holm adjustment ($P_{\text{adj}} < 0.0001$). In Table 3, the observation-only and schema-only variants reach macro-AUPRC values of 0.794 (95% CI: 0.781–0.816) and 0.800 (0.785–0.821), respectively, compared with 0.545 (0.538–0.572) for the Baseline. Combining the two components yields a macro-AUPRC of 0.802 (0.788–0.823) and the highest precision.

Table 3: Component ablation of AMA. Values are estimates with 95% bootstrap confidence intervals. Statistical comparisons use the Baseline as the reference.

Method	Precision^{Mi} (95% CI)	P_{adj}	Specificity^{Mi} (95% CI)	P_{adj}	F1^{Ma} (95% CI)	P_{adj}	Bal. Acc.^{Mi} (95% CI)	P_{adj}	κ^{Ma} (95% CI)	P_{adj}	AUPRC^{Ma} (95% CI)	P_{adj}
Baseline	0.704 (0.700–0.708)	Ref.	0.930 (0.929–0.931)	Ref.	0.495 (0.478–0.508)	Ref.	0.874 (0.872–0.876)	Ref.	0.423 (0.404–0.437)	Ref.	0.545 (0.538–0.572)	Ref.
Observation-only	0.875 (0.872–0.878)	****	0.974 (0.974–0.975)	****	0.751 (0.728–0.764)	****	**0.935 (0.933–0.936)**	****	0.714 (0.689–0.727)	****	0.794 (0.781–0.816)	****
Schema-only	0.890 (0.887–0.893)	****	**0.978 (0.977–0.978)**	****	**0.759 (0.734–0.774)**	****	0.932 (0.931–0.934)	****	**0.723 (0.697–0.739)**	****	0.800 (0.785–0.821)	****
AMA	**0.892 (0.889–0.895)**	****	**0.978 (0.978–0.979)**	****	0.755 (0.730–0.769)	****	**0.935 (0.934–0.937)**	****	0.718 (0.693–0.733)	****	**0.802 (0.788–0.823)**	****

^{Mi} Micro aggregation; ^{Ma} macro aggregation. Holm-adjusted significance: ***$P < 0.001$ and ****$P < 0.0001$. Bold and underline indicate the best and second-best estimates, respectively; ties after rounding retain their source-table formatting.

Hyperparameter Sensitivity. We vary the observation-loss coefficient λ_{obs} and the schema-balancing exponent γ independently while holding the encoder, data split, and all remaining optimization settings fixed. The primary configuration uses $\lambda_{\text{obs}} = 0.5$ and $\gamma = 0.5$. Figure S2 shows that macro-AUPRC remains within a narrow range across the evaluated settings, with the primary setting $\lambda_{\text{obs}} = \gamma = 0.5$ attaining the highest estimate. Table S4 reports the corresponding six evaluation measures and bootstrap intervals.

4.4 Robustness and Analysis

Cross-backbone Evaluation. AMA improves macro-F1 and macro-averaged Cohen's κ across RETFound-MAE-MEH, RETFound-DINOv2-MEH, RetiZero, and RetFiner, spanning three pre-training paradigms (Figure 5; all $P_{\text{adj}} < 0.0001$). Paired bootstrap mean macro-F1 gains range from 0.122 to 0.258 and macro-κ gains from 0.138 to 0.294. Table S5 provides confidence intervals and

additional metrics, including positive macro-AUPRC gains for every backbone and calibration results across encoders (Guo et al., 2017).

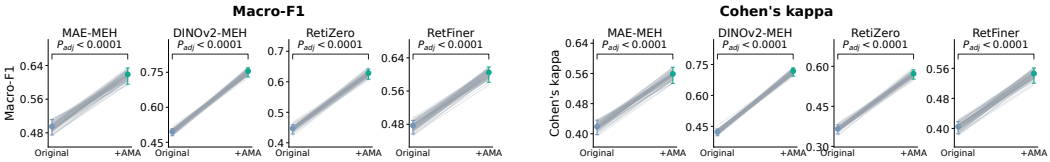

Figure 5: Cross-backbone comparison of AMA across four pre-trained retinal backbones. The four left panels report macro-F1 and the four right panels report macro-averaged Cohen's κ, each ordered as RETFound-MAE-MEH, RETFound-DINOv2-MEH, RetiZero, and RetFiner. Large points and error bars denote estimates and 95% bootstrap confidence intervals; thin lines connect aligned bootstrap replicates of each original model and its corresponding AMA model. Reported P values are Holm-adjusted paired comparisons.

Label-level Analysis. Figure 6 examines whether the aggregate gain is broadly distributed across labels. AMA improves the AUPRC point estimate for 96 labels, with paired 95% confidence intervals entirely above zero for 73 labels. Additional category-stratified and support-dependent analyses are reported in Figure S4.

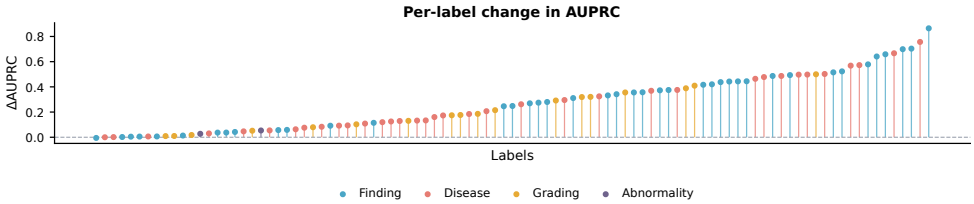

Figure 6: Label-level analysis of AMA. Sorted per-label AUPRC changes relative to the matched Baseline.

5 Conclusion

AMA addresses multi-source retinal learning under heterogeneous annotation mechanisms through an auxiliary label-observability objective and source–label schema balancing. On the unified 98-label task, AMA attains the highest macro-AUPRC point estimate among the evaluated multi-label objectives under the primary encoder, while its gains over masked supervision persist across four backbones. The present evaluation focuses on public fundus datasets and a curated label mapping; future work should examine prospective cohorts, additional ophthalmic modalities, and calibration across backbones.

AI Use Statement. Generative AI assisted with literature review, study planning, development of analysis and visualization code, figure prototyping, manuscript organization, and language editing. It generated neither experimental data nor clinical annotations. The authors verified all sources, code, analyses, and results and take responsibility for the paper. **Ethics Statement.** We used public retinal datasets in accordance with their terms and made no attempt at re-identification. Models may inherit population and annotation biases; these results do not establish safety for autonomous clinical use, which requires prospective validation and clinical oversight. **Reproducibility Statement.** The main text specifies the formulation and evaluation protocol; the appendix provides dataset and label documentation and additional analyses. Code, experiment configurations, and evaluation scripts will be released upon publication.

REFERENCES

Emanuel Ben-Baruch, Tal Ridnik, Itamar Friedman, Avi Ben-Cohen, Nadav Zamir, Asaf Noy, and Lihi Zelnik-Manor. Multi-label classification with partial annotations using class-aware selective loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4764–4772, 2022.

Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 933–942, 2021.

Coen de Vente, Koenraad A. Vermeer, Nicolas Jaccard, et al. AIROGS: Artificial intelligence for robust glaucoma screening challenge. *IEEE Transactions on Medical Imaging*, 2024. doi: 10.1109/TMI.2023.3313786. URL https://doi.org/10.1109/TMI.2023.3313786.

Jiawei Du, Jia Guo, Weihang Zhang, Shengzhu Yang, Hanruo Liu, Huiqi Li, and Ningli Wang. RET-CLIP: A retinal image foundation model pre-trained with clinical diagnostic reports. In *Medical Image Computing and Computer Assisted Intervention*, pp. 709–719, 2024.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

Ronald Fecso, José Morano, Ursula Schmidt-Erfurth, and Hrvoje Bogunović. RetFiner: A vision-language refinement scheme for retinal foundation models. In *Medical Image Computing and Computer Assisted Intervention*, pp. 543–553, 2025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Youngwook Kim, Jae Myung Kim, Zeynep Akata, and Jungwoo Lee. Large loss matters in weakly supervised multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14156–14165, 2022.

Yeonkyung Lee, Woojung Han, Youngjun Jun, Hyeonmin Kim, Jungkyung Cho, and Seong Jae Hwang. PRETI: Patient-aware retinal foundation model via metadata-guided representation learning. In *Medical Image Computing and Computer Assisted Intervention*, 2025.

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*, 2025.

Zihan Li, Yiqing Wang, Sina Farsiu, and Paul Kinahan. Boosting medical visual understanding from multi-granular language learning. In *International Conference on Learning Representations*, 2026.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

José Morano, Botond Fazekas, Emese Sükei, Ronald Fecso, Taha Emre, Markus Gumpinger, Georg Faustmann, Marzieh Oghbaie, Ursula Schmidt-Erfurth, and Hrvoje Bogunović. MIRAGE: Multimodal foundation model and benchmark for comprehensive retinal OCT image analysis. *npj Digital Medicine*, 2025. doi: 10.1038/s41746-025-01852-3.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Jianing Qiu, Jian Wu, Hao Wei, Peilun Shi, Minqing Zhang, Yunyun Sun, Lin Li, Hanruo Liu, Hongyi Liu, Simeng Hou, et al. Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence. *NEJM AI*, 1(12):AIoa2300221, 2024. doi: 10.1056/AIoa2300221.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 82–91, 2021.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: 10.1093/biomet/63.3.581.

Takaya Saito and Marc Rehmsmeier. The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015. doi: 10.1371/journal.pone.0118432.

Danli Shi, Weiyi Zhang, Jiancheng Yang, Siyu Huang, Xiaolan Chen, Mayinuer Yusufu, Kai Jin, Shan Lin, Shunming Liu, Qing Zhang, et al. A multimodal visual–language foundation model for computational ophthalmology. *npj Digital Medicine*, 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*, 2024. doi: 10.1016/j.media.2024.103357.

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

Ruiqi Wu, Chenran Zhang, Jianle Zhang, Yi Zhou, Tao Zhou, and Huazhu Fu. MM-Retinal: Knowledge-enhanced foundational pretraining with fundus image-text expertise. In *Medical Image Computing and Computer Assisted Intervention*, 2024.

Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European Conference on Computer Vision*, pp. 162–178, 2020.

Shengzhu Yang, Jiawei Du, Jia Guo, Weihang Zhang, Hanruo Liu, Huiqi Li, and Ningli Wang. ViLReF: An expert knowledge enabled vision-language retinal foundation model, 2024.

Kai Yu, Yang Zhou, Yang Bai, Zhi Da Soh, Xinxing Xu, Rick Siow Mong Goh, Ching-Yu Cheng, and Yong Liu. UrFound: Towards universal retinal foundation models via knowledge-guided masked modeling. In *Medical Image Computing and Computer Assisted Intervention*, pp. 746–756, 2024.

Donghao Zhou, Pengfei Chen, Qiong Wang, Guangyong Chen, and Pheng-Ann Heng. Acknowledging the unknown for multi-label learning with single positive labels. In *European Conference on Computer Vision*, pp. 423–440, 2022.

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*, 622:156–163, 2023. doi: 10.1038/s41586-023-06555-x.

Supplementary Material

Learning from Heterogeneous Annotation Mechanisms for Unified Retinal Prediction

A Dataset Documentation

A.1 Data Sources and Splits

The unified corpus contains 252,116 retinal images collected from 54 public datasets. We retain source identity throughout harmonization because label observability is strongly structured by source-specific annotation protocols. Table S1 reports the source inventory, the subset of the unified vocabulary covered by each source, and the fixed training, validation, and test split. Dataset-level references are listed in the Dataset References section at the end of this appendix.

Table S1: Dataset inventory and unified-label coverage. Label identifiers refer to Table S2. Split counts are image counts after harmonization and quality control.

ID	Dataset	Covered label IDs	Train	Val.	Test	Total
1	ACRIMA705[D1]	24, 89	483	69	153	705
2	ADAM400[D2]	10, 89	855	109	232	1,196
3	AIROGS[D3]	89	71,060	10,231	20,151	101,442
4	AOD[D4]	10, 16, 21, 24, 30, 89	10,363	1,495	2,955	14,813
5	APTOS3662[D5]	3–9, 21, 89	2,455	354	725	3,534
6	ARIA143[D6]	10, 89	99	9	35	143
7	BEH[D7]	24, 89	443	63	128	634
8	BRSET[D8]	3–10, 21, 26, 28, 37, 39, 54, 60, 75–77, 89	11,354	1,640	3,272	16,266
9	BiDR[D9]	21, 89	2,011	269	558	2,838
10	BoVW[D10]	52, 54, 56, 59, 75, 80–81, 89	1,426	202	385	2,013
11	CHAKSU[D11]	24, 87	892	145	301	1,338
12	Cataract[D12]	16, 24, 89	388	65	148	601
13	DDR[D13]	3–9, 21, 56, 59, 65, 75, 80, 89	9,531	1,361	2,687	13,579
14	DGOCF[D14]	3–9, 21, 89	6,906	1,013	2,020	9,939
15	DHRF[D15]	16, 21, 24, 27, 39, 67, 89	1,890	281	574	2,745
16	DeepDRiD1600[D16]	3–9, 21, 89	1,389	201	410	2,000
17	DiaRetDB[D17]	89	65	6	18	89
18	E-ophta[D18]	56, 65, 89	297	45	92	434
19	EyePACS35126[D19]	3–9, 21, 89	24,684	3,462	6,980	35,126
20	FIVES[D20]	10, 21, 24, 89	562	71	167	800
21	FUND[D21]	10, 17, 24, 28, 89	131	20	28	179
22	G1020[D22]	24, 89	731	91	198	1,020
23	GAMMA100[D23]	24, 89–92	69	11	20	100
24	GNG[D24]	24, 89	295	30	75	400
25	HRF45[D25]	21, 24, 89	38	3	4	45
26	HarvardGlaucoma[D26]	24, 89	1,081	154	305	1,540
27	IDRID516[D27]	0–9, 20–21, 28, 89	360	53	103	516
28	INSPIRE-AVR40[D28]	89	33	1	6	40
29	IOSTAR30[D29]	89	22	5	3	30
30	JSIEC[D30]	13, 15, 17, 19, 21, 23, 26, 29–30, 33, 36–37, 39–42, 44, 51, 53, 55–56, 58–61, 63–64, 66, 68, 70, 74–75, 80, 82, 85–89, 93–97	685	105	206	996
31	LAG[D31]	87	3,407	463	984	4,854
32	LSD[D32]	61, 89	125	17	33	175
33	LongDRScreening-2520[D33]	89	763	118	239	1,120
34	MESSIDOR2[D34]	3–9, 20–21, 28, 89	1,201	178	369	1,748
35	MMAC_TASK1[D35]	93–97	797	112	227	1,136
36	MMAMD_CFP[D36]	10, 22, 34, 45, 89	764	110	217	1,091
37	MuReD[D37]	10, 12, 15, 17, 19, 21, 26, 30, 33, 39–40, 48–49, 54, 61, 63, 69–71, 78, 89, 93–97	1,729	230	492	2,451
38	ODIR[D38]	10, 16, 24, 30, 33, 89	4,885	691	1,424	7,000
39	ORIGA650[D39]	24, 89	475	54	121	650
40	PALM1200[D40]	30, 33, 89	831	110	243	1,184
41	PAPILA[D41]	24, 87, 89	335	56	97	488
42	PARAGUAY[D42]	3–9, 21, 89	524	81	152	757
43	Papilledema[D43]	32, 35, 89	961	138	270	1,369

Continued on next page

Table S1 (continued)

ID	Dataset	Covered label IDs	Train	Val.	Test	Total
44	REFUGE1200[D44]	24, 89	841	117	242	1,200
45	RFMiD2[D45]	10–11, 15, 17–19, 21, 25, 30, 33, 36–41, 43, 46–48, 50, 54–57, 61–63, 69–72, 74–75, 78–80, 82–83, 89, 93–97	591	91	175	857
46	RFMiD3200[D46]	10–11, 14–15, 17–19, 21, 25, 30, 33, 36–41, 43, 46–48, 50, 54–57, 61–63, 66, 69–75, 78–80, 82–84, 89, 93–97	2,205	337	658	3,200
47	RIM-ONE[D47]	24, 89	348	50	87	485
48	RITE40[D48]	89	29	2	9	40
49	ROD[D49]	15, 19, 36, 39–40, 89	195	30	56	281
50	ROI[D50]	21, 89	781	104	235	1,120
51	STARE[D51]	89	276	49	72	397
52	SUSTech-SYSU[D52]	3–9, 21, 89	874	113	232	1,219
53	ToxoFundus[D53]	31, 89	288	36	87	411
54	VietAI[D54]	21, 24, 27–28, 39, 67, 89	2,658	361	763	3,782
	Total		**176,481**	**25,212**	**50,423**	**252,116**

A.2 Unified Label Vocabulary and Training Support

The 98-label vocabulary comprises one abnormality label, 37 disease labels, 18 grading labels, and 42 retinal finding labels. Table S2 reports training-set positive support and the number of training sources in which each label is observed. A source is counted as observing a label when it contains at least one positive or confirmed negative entry for that label after harmonization. This definition characterizes the supervision available during learning and does not use test-set prevalence or metric eligibility.

Table S2: Unified 98-label dictionary and training support. Observing sources count training sources containing at least one observed positive or negative entry for a label after label harmonization. Counts use training data and exclude unobserved entries and test-set statistics.

ID	Canonical label	Train positives	Observing sources
Abnormality			
89	Fundus abnormality status	48,005	51
Disease			
10	Age-related macular degeneration (AMD)	3,097	11
11	Anterior ischemic optic neuropathy	21	2
12	Arteriosclerotic retinopathy	23	1
13	Bietti crystalline dystrophy	6	1
14	Branch retinal artery occlusion	2	1
15	Branch retinal vein occlusion	291	5
16	Cataract	2,030	4
17	Central serous chorioretinopathy (CSCR/CSR)	94	5
18	Central retinal artery occlusion	3	2
19	Central retinal vein occlusion	141	5
20	Diabetic macular edema (DME)	309	2
21	Diabetic retinopathy (DR)	22,392	21
22	Dry age-related macular degeneration	74	1
23	Fundus neoplasm	6	1
24	Glaucoma	4,300	19
25	Hemorrhagic retinopathy	60	2
26	Hypertensive retinopathy	232	3
27	Macular degeneration (generic)	777	2
28	Macular edema (disease)	965	5
29	Maculopathy (generic)	52	1
30	Myopia (generic)	2,170	7
31	Ocular toxoplasmosis	195	1

Continued on next page

Table S2 (continued)

ID	Canonical label	Train positives	Observing sources
32	Papilledema	207	1
33	Pathological myopia	871	6
34	Polypoidal choroidal vasculopathy	209	1
35	Pseudopapilledema	206	1
36	Retinal artery occlusion (RAO, parent)	34	4
37	Retinal detachment (RD)	62	4
38	Retinal telangiectasia	45	2
39	Retinal vascular occlusion (parent)	994	8
40	Retinal vein occlusion (RVO, parent)	432	5
41	Retinitis pigmentosa	26	3
42	Rhegmatogenous retinal detachment	40	1
43	Strabismus	10	2
44	Vogt-Koyanagi-Harada disease	10	1
45	Wet age-related macular degeneration	344	1
87	Glaucoma suspect	1,399	4
Grading			
0	No DME	152	1
1	CSME	35	1
2	Severe CSME	168	1
3	No DR	40,617	10
4	Mild NPDR	4,431	10
5	Moderate NPDR	8,662	10
6	Severe NPDR	1,585	10
7	Very Severe NPDR	75	10
8	PDR	2,445	10
9	Advanced PDR	81	10
90	Non-glaucoma	35	1
91	Early glaucoma	18	1
92	Mid-advanced glaucoma	16	1
93	No macular lesions	279	5
94	Tessellated fundus	646	5
95	Diffuse chorioretinal atrophy	157	5
96	Patchy chorioretinal atrophy	42	5
97	Macular atrophy	30	5
Finding			
46	Asteroid hyalosis	19	2
47	Chorioretinal folds	9	2
48	Chorioretinal scar	88	3
49	Choroidal neovascularization	42	1
50	Collateral vessels (coarse)	5	2
51	Congenital disc abnormality	7	1
52	Deep hemorrhage	102	1
53	Dragged disc	7	1
54	Drusen	2,375	5
55	Epiretinal membrane	40	3
56	Exudate (generic)	760	6
57	Exudation (generic)	66	2
58	Fibrosis	7	1
59	Hard exudate	570	3
60	Increased cup-disc ratio	2,276	2
61	Laser spot/scar	235	5
62	Large macular hemorrhage	16	2
63	Macular hole	529	4
64	Massive hard exudates	9	1
65	Microaneurysm	483	2
66	Myelinated nerve fibers	10	2

Continued on next page

Table S2 (continued)

ID	Canonical label	Train positives	Observing sources
67	Ocular opacity (corneal/lens/vitreal)	1,345	2
68	Optic atrophy	9	1
69	Optic disc cupping	522	3
70	Optic disc edema / disc swelling	130	4
71	Optic disc pallor	136	3
72	Optic disc pigmentation	2	2
73	Retinal plaque	2	1
74	Preretinal hemorrhage	20	3
75	Retinal hemorrhage	897	6
76	Retinal nevus	93	1
77	Retinal scar	203	1
78	Retinoschisis	95	3
79	RPE changes	26	2
80	Soft exudate / cotton-wool spot	264	5
81	Superficial hemorrhage	72	1
82	Tortuous vessels	30	3
83	Vascular sheathing	7	2
84	Vitreous hemorrhage	3	1
85	Vitreous particles	9	1
86	Yellow-white spots / flecks	20	1
88	Silicone oil in eye	13	1

B Evaluation Protocol

Model selection and held-out evaluation. All primary models are optimized for 100 epochs with AdamW. The checkpoint with the highest validation macro-AUPRC is selected, and its operating thresholds are determined on the validation split. The selected checkpoint and fixed thresholds are then applied once to the held-out test split. Unless the encoder is the experimental variable, AMA and its comparator share the same encoder, data partition, optimization schedule, and training budget.

Uncertainty and paired comparisons. Statistics are computed only on observed test entries. Macro metrics average over labels with at least five observed positives and five observed negatives; micro metrics pool all observed entries. Confidence intervals use 500 ordinary prediction-row bootstrap resamples. For label-wise and macro quantities, observed prediction rows are resampled within each label before aggregation; for micro quantities, observed rows are pooled before resampling. These intervals quantify test-sample uncertainty rather than variation across retraining runs. Comparisons preserve aligned resamples for each method pair. Two-sided paired P values use a normal approximation based on the bootstrap mean and standard error of paired differences. Holm adjustment is applied jointly to all paired tests across the six reported measures within each analysis table. Reported paired changes are bootstrap mean differences and may differ from differences between full-sample point estimates. All source–label support statistics, class weights, and schema-balancing coefficients are computed from the training partition only.

C Additional Experimental Results

C.1 Comparison of Supervision Variants

Table S3 contrasts three supervision policies on identical observed test entries. AMA attains the highest point estimates across all six measures, followed by the `U`-as-negative variant and the masked Baseline. Because this supplementary comparison was not designed as a single-factor intervention, it is interpreted as a comparison of complete training policies rather than an estimate of the isolated effect of relabeling `U`.

Table S3: Comparison of supervision variants using RETFound-DINOv2-MEH. Values are estimates with 95% bootstrap confidence intervals from 500 prediction-row resamples; inferential comparisons are not performed for this descriptive analysis. All variants are evaluated on identical observed image–label pairs; macro measures use the same label-eligibility criterion as the main experiments.

Method	Precision^{Mi} (95% CI)	Specificity^{Mi} (95% CI)	F1^{Ma} (95% CI)	BAcc^{Mi} (95% CI)	κ^{Ma} (95% CI)	AUPRC^{Ma} (95% CI)
U-as-negative	0.864 (0.861–0.867)	0.972 (0.972–0.973)	0.713 (0.692–0.727)	0.921 (0.919–0.922)	0.663 (0.642–0.678)	0.758 (0.748–0.781)
Masked Baseline	0.704 (0.700–0.708)	0.930 (0.929–0.931)	0.495 (0.478–0.508)	0.874 (0.872–0.876)	0.423 (0.404–0.437)	0.545 (0.538–0.572)
AMA	**0.892 (0.889–0.895)**	**0.978 (0.978–0.979)**	**0.755 (0.730–0.769)**	**0.935 (0.934–0.937)**	**0.718 (0.693–0.733)**	**0.802 (0.788–0.823)**

C.2 Component Ablation

Figure S1 complements the component comparison in the main text.

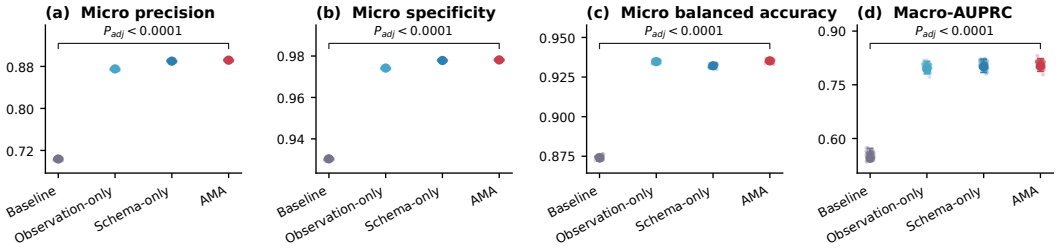

Figure S1: Component ablation of AMA for micro precision, micro specificity, micro balanced accuracy, and macro-AUPRC. Large points and error bars denote estimates and 95% bootstrap confidence intervals; small points show aligned bootstrap replicates. Plot annotations report Holm-adjusted paired comparisons with the Baseline.

C.3 Hyperparameter Sensitivity

We vary the observation-loss coefficient λ_{obs} and the schema-balancing exponent γ independently while holding the encoder, data split, and remaining optimization settings fixed. The primary configuration $\lambda_{\text{obs}} = \gamma = 0.5$ yields a macro-AUPRC of 0.802; estimates range from 0.788 to 0.802 across the evaluated settings.

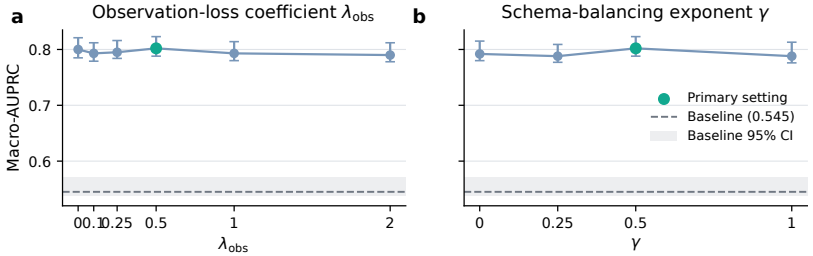

Figure S2: Hyperparameter sensitivity of AMA. Macro-AUPRC is evaluated while varying **(a)** the observation-loss coefficient and **(b)** the schema-balancing exponent. Points and error bars denote estimates and 95% bootstrap confidence intervals; green markers identify the primary setting. The dashed horizontal line and gray band denote the matched Baseline macro-AUPRC (0.545; 95% CI: 0.538–0.572).

Table S4: Hyperparameter sensitivity of AMA. Values are estimates with 95% bootstrap confidence intervals. The primary setting is shown in bold.

Parameter	Value	Precision^{Mi} (95% CI)	Specificity^{Mi} (95% CI)	F1^{Ma} (95% CI)	Bal. Acc.^{Mi} (95% CI)	κ^{Ma} (95% CI)	AUPRC^{Ma} (95% CI)
λ_{obs}	0	0.890 (0.887–0.893)	0.978 (0.977–0.978)	0.759 (0.734–0.774)	0.932 (0.931–0.934)	0.723 (0.697–0.739)	0.800 (0.785–0.821)
	0.10	0.891 (0.888–0.893)	0.978 (0.977–0.979)	0.755 (0.732–0.769)	0.932 (0.931–0.934)	0.719 (0.695–0.735)	0.793 (0.779–0.812)
	0.25	0.874 (0.871–0.877)	0.974 (0.973–0.974)	0.760 (0.738–0.776)	0.937 (0.936–0.939)	0.723 (0.701–0.739)	0.795 (0.784–0.816)
	0.50	**0.892 (0.889–0.895)**	**0.978 (0.978–0.979)**	**0.755 (0.730–0.769)**	**0.935 (0.934–0.937)**	**0.718 (0.693–0.733)**	**0.802 (0.788–0.823)**
	1.00	0.868 (0.865–0.871)	0.972 (0.972–0.973)	0.745 (0.725–0.758)	0.935 (0.934–0.937)	0.706 (0.685–0.719)	0.793 (0.780–0.814)
	2.00	0.880 (0.877–0.883)	0.975 (0.975–0.976)	0.748 (0.723–0.761)	0.934 (0.932–0.935)	0.709 (0.684–0.724)	0.790 (0.778–0.812)
γ	0	0.880 (0.877–0.883)	0.975 (0.975–0.976)	0.750 (0.729–0.761)	0.934 (0.933–0.936)	0.708 (0.686–0.720)	0.792 (0.780–0.815)
	0.25	0.877 (0.874–0.880)	0.975 (0.974–0.975)	0.735 (0.712–0.749)	0.934 (0.932–0.935)	0.696 (0.672–0.711)	0.788 (0.777–0.809)
	0.50	**0.892 (0.889–0.895)**	**0.978 (0.978–0.979)**	**0.755 (0.730–0.769)**	**0.935 (0.934–0.937)**	**0.718 (0.693–0.733)**	**0.802 (0.788–0.823)**
	1.00	0.878 (0.876–0.882)	0.975 (0.974–0.975)	0.754 (0.733–0.767)	0.937 (0.935–0.938)	0.716 (0.694–0.731)	0.788 (0.776–0.813)

CIs use 500 prediction-row bootstrap replicates. The $\lambda_{\text{obs}} = 0$ row corresponds to the schema-only variant; $\lambda_{\text{obs}} = \gamma = 0.5$ is the primary setting.

C.4 Cross-backbone Evaluation Metrics

Table S5 complements the paired F1 and Cohen's kappa comparisons in the main paper with six evaluation measures for each original backbone and its corresponding AMA model.

Table S5: Cross-backbone evaluation of AMA. Values are estimates with 95% bootstrap confidence intervals. Each original model is the reference for its corresponding AMA model.

Method	Sens.^{Mi} (95% CI)	P_{adj}	ECE^{Mi} (95% CI)	P_{adj}	F1^{Ma} (95% CI)	P_{adj}	Bal. Acc.^{Mi} (95% CI)	P_{adj}	κ^{Ma} (95% CI)	P_{adj}	AUPRC^{Ma} (95% CI)	P_{adj}
RETFound-MAE-MEH	0.785 (0.781–0.788)	Ref.	**0.034 (0.033–0.035)**	Ref.	0.495 (0.475–0.511)	Ref.	0.854 (0.852–0.856)	Ref.	0.419 (0.398–0.436)	Ref.	0.529 (0.520–0.555)	Ref.
+ AMA	**0.821 (0.817–0.825)**	****	0.054 (0.053–0.054)	****	**0.619 (0.596–0.634)**	****	**0.889 (0.886–0.890)**	****	**0.559 (0.534–0.577)**	****	**0.648 (0.639–0.676)**	****
RETFound-DINOv2-MEH	0.818 (0.814–0.821)	Ref.	**0.019 (0.018–0.020)**	Ref.	0.495 (0.478–0.508)	Ref.	0.874 (0.872–0.876)	Ref.	0.423 (0.404–0.437)	Ref.	0.545 (0.538–0.572)	Ref.
+ AMA	**0.892 (0.889–0.895)**	****	0.032 (0.032–0.033)	****	**0.755 (0.730–0.769)**	****	**0.935 (0.934–0.937)**	****	**0.718 (0.693–0.733)**	****	**0.802 (0.788–0.823)**	****
RetiZero	0.762 (0.758–0.765)	Ref.	0.065 (0.064–0.066)	Ref.	0.447 (0.429–0.459)	Ref.	0.851 (0.849–0.853)	Ref.	0.366 (0.346–0.380)	Ref.	0.446 (0.441–0.471)	Ref.
+ AMA	**0.836 (0.833–0.840)**	****	**0.045 (0.045–0.046)**	****	**0.628 (0.608–0.642)**	****	**0.899 (0.897–0.901)**	****	**0.572 (0.552–0.587)**	****	**0.649 (0.641–0.675)**	****
RetFiner	0.804 (0.801–0.808)	Ref.	**0.014 (0.014–0.015)**	Ref.	0.477 (0.457–0.489)	Ref.	0.862 (0.860–0.864)	Ref.	0.405 (0.386–0.419)	Ref.	0.524 (0.515–0.551)	Ref.
+ AMA	**0.811 (0.808–0.815)**	–	0.058 (0.057–0.059)	****	**0.605 (0.581–0.618)**	****	**0.883 (0.881–0.885)**	****	**0.546 (0.520–0.561)**	****	**0.640 (0.630–0.665)**	****

^{Mi} Micro aggregation; ^{Ma} macro aggregation. Lower ECE is better. Holm-adjusted significance: $^{****}P < 0.0001$; "–" does not meet this threshold. Bold indicates the better estimate within each backbone pair.

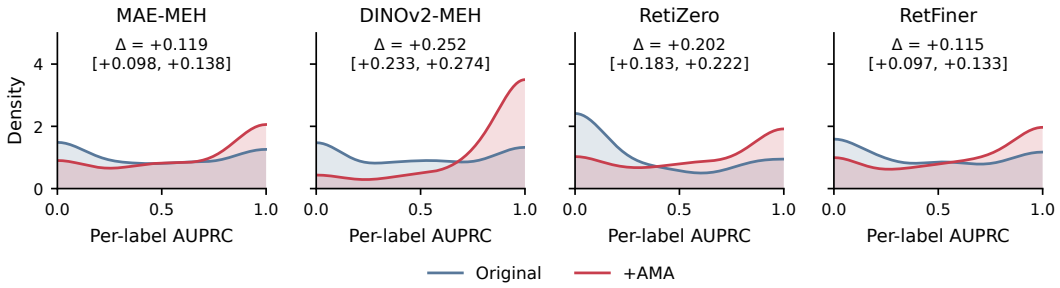

Figure S3: Per-label AUPRC distributions for the original and corresponding AMA models across four backbones. MAE-MEH and DINOv2-MEH denote the corresponding RETFound models. Distributions use labels with finite AUPRC for both models and a common boundary-aware kernel density estimate within each pair. Insets report paired-bootstrap mean changes in macro-AUPRC and their 95% confidence intervals.

C.5 Additional Label-level Analyses

The gain in per-label AUPRC is negatively associated with observed positive training support (Spearman $\rho = -0.36$, bootstrap 95% CI: -0.54 to -0.15; $P < 0.001$). Improvements are present across all four annotation categories and are largest for retinal findings.

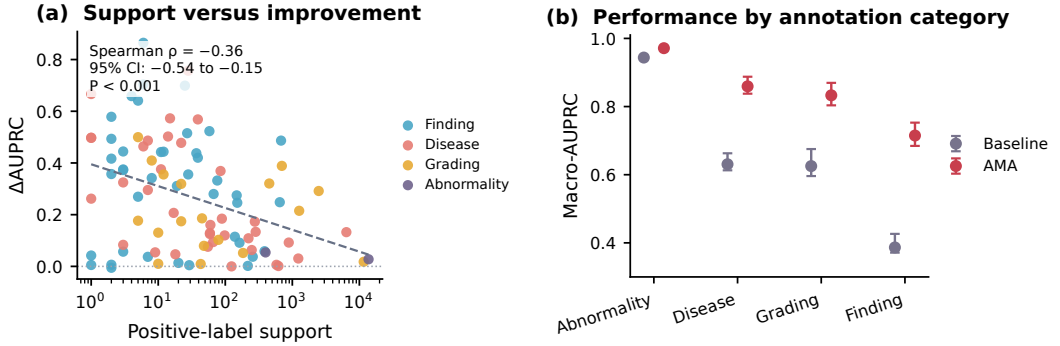

Figure S4: Additional label-level analyses. **Left:** positive-label support versus AUPRC improvement; the Spearman interval uses 10,000 label-level bootstrap samples. **Right:** Baseline and AMA macro-AUPRC across annotation categories, with 95% bootstrap confidence intervals.

D REPRESENTATION ANALYSIS

Figure S5 provides a qualitative comparison of feature geometry across four representation settings.

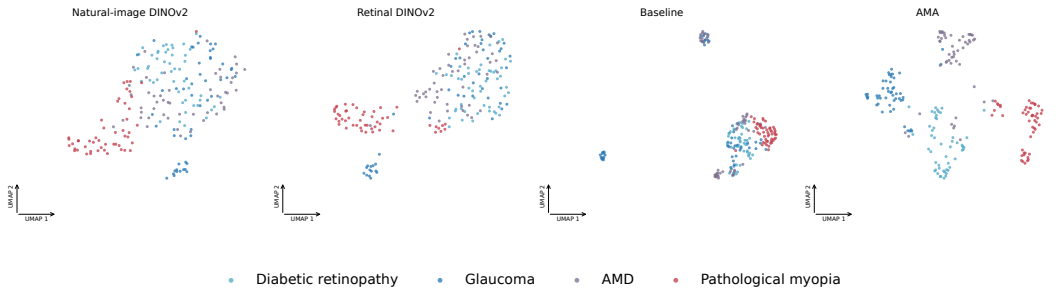

Figure S5: UMAP visualization of retinal image representations from a natural-image DINOv2 encoder, a retinal DINOv2 encoder, the matched Baseline, and AMA. Points are colored by four representative diagnostic groups: diabetic retinopathy, glaucoma, age-related macular degeneration (AMD), and pathological myopia.

DATASET REFERENCES

The following references document the 54 constituent datasets listed in Table S1. Dataset-reference identifiers are kept separate from the main bibliography.

[D1] ACRIMA705. Deep Learning & Digital Fundus Images: Glaucoma Detection using DenseNet. https://figshare.com/s/c2d31f850af14c5b5232

[D2] ADAM400. ADAM Challenge: Detecting Age-Related Macular Degeneration From Fundus Images. https://ieee-dataport.org/documents/adam-automatic-detection-challenge-age-related-macular-degeneration

[D3] AIROGS. AIROGS: Artificial Intelligence for Robust Glaucoma Screening Challenge. https://airogs.grand-challenge.org/data-and-challenge/

[D4] AOD. Augmented Ocular Diseases. https://www.kaggle.com/datasets/nurmukhammed7/augemnted-ocular-diseases

[D5] APTOS3662. APTOS 2019 Blindness Detection, Kaggle Competition. https://www.kaggle.com/competitions/aptos2019-blindness-detection

[D6] ARIA143. Automated Retinal Image Analysis (ARIA) dataset. http://www.eyecharity.com/aria_online/

[D7] BEH. Deep Learning-Based Glaucoma Detection With Cropped Optic Cup and Disc and Blood Vessel Segmentation. https://www.kaggle.com/datasets/kabirjaan/beh-glaucoma

[D8] BRSET. BRSET: A Brazilian Multilabel Ophthalmological Dataset of Retina Fundus Photos. https://physionet.org/content/brazilian-ophthalmological/1.0.1/

[D9] BiDR. Diagnosis of Diabetic Retinopathy. https://www.kaggle.com/datasets/pkdarabi/diagnosis-of-diabetic-retinopathy

[D10] BoVW. Advancing Bag-of-Visual-Words Representations for Lesion Classification in Retinal Images. https://doi.org/10.6084/m9.figshare.953671

[D11] CHAKSU. Chákṣu: A glaucoma specific fundus image database. https://doi.org/10.6084/m9.figshare.20123135

[D12] Cataract. Cataract dataset. https://www.kaggle.com/datasets/jr2ngb/cataractdataset

[D13] DDR. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. https://github.com/nkicsl/DDR-dataset

[D14] DGOCF. Applying artificial intelligence to disease staging: Deep learning for improved staging of diabetic retinopathy. https://doi.org/10.6084/m9.figshare.4879853

[D15] DHRF. Derbi hackathon retinal fundus image dataset. https://www.kaggle.com/datasets/nikkich9/derbi-hackathon-retinal-fundus-image-dataset

[D16] DeepDRiD1600. DeepDRiD: Diabetic Retinopathy—Grading and Image Quality Estimation Challenge. https://zenodo.org/records/6452623

[D17] DiaRetDB. The DIARETDB1 diabetic retinopathy database and evaluation protocol. https://www.kaggle.com/datasets/nguyenhung1903/diaretdb1-v21

[D18] E-ophta. TeleOphta: Machine learning and image processing methods for teleophthalmology. https://www.adcis.net/en/third-party/e-ophtha/

[D19] EyePACS35126. EyePACS/Kaggle Diabetic Retinopathy Detection Challenge. https://www.kaggle.com/c/diabetic-retinopathy-detection

[D20] FIVES. FIVES: A Fundus Image Dataset for Artificial Intelligence based Vessel Segmentation. https://doi.org/10.6084/m9.figshare.19688169

[D21] FUND. A Composite Retinal Fundus and OCT Dataset to Grade Macular and Glaucomatous Disorders. https://data.mendeley.com/datasets/trghs22fpg/4

[D22] G1020. G1020: A Benchmark Retinal Fundus Image Dataset for Computer-Aided Glaucoma Detection. https://www.kaggle.com/datasets/kiamahmed/glaucoma-fundus-imaging-g1020-splitted

[D23] GAMMA100. GAMMA challenge: Glaucoma grAding from Multi-Modality imAges. https://huggingface.co/datasets/ctmedtech/GAMMA

[D24] GNG. REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. https://refuge.grand-challenge.org/Details/

[D25] HRF45. Robust Vessel Segmentation in Fundus Images. https://www5.cs.fau.de/research/data/fundus-images/

[D26] HarvardGlaucoma. Harvard Glaucoma Detection and Progression: A Multimodal Multitask Dataset and Generalization-Reinforced Semi-Supervised Learning. https://github.com/Harvard-Ophthalmology-AI-Lab/Harvard-GDP

[D27] IDRID516. Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research. https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid

[D28] INSPIRE-AVR40. Automated Measurement of the Arteriolar-To-Venular Width Ratio in Digital Color Fundus Photographs. https://eye.medicine.uiowa.edu/inspire-datasets

[D29] IOSTAR30. Robust Retinal Vessel Segmentation via Locally Adaptive Derivative Frames in Orientation Scores. https://www.retinacheck.org/download-iostar-retinal-vessel-segmentation-dataset

[D30] JSIEC. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. https://www.kaggle.com/datasets/linchundan/fundusimage1000

[D31] LAG. Attention Based Glaucoma Detection: A Large-Scale Database and CNN Model. https://github.com/smilell/AG-CNN

[D32] LSD. Laser Scar Detection in Fundus Images Using Convolutional Neural Networks. https://github.com/li-xirong/fundus10k/

[D33] LongDRScreening2520. Longitudinal diabetic retinopathy screening data. https://academictorrents.com/details/744717095e59373186abec814c86de4831d889e9

[D34] MESSIDOR2. Automated Analysis of Retinal Images for Detection of Referable Diabetic Retinopathy. https://www.adcis.net/en/third-party/messidor2/

[D35] MMAC_TASK1. A competition for the diagnosis of myopic maculopathy by artificial intelligence algorithms. https://codalab.lisn.upsaclay.fr/competitions/12477

[D36] MMAMD_CFP. Learning Two-Stream CNN for Multi-Modal Age-related Macular Degeneration Categorization. https://github.com/li-xirong/mmc-amd

[D37] MuReD. Multi-Label Retinal Disease Classification Using Transformers. https://data.mendeley.com/datasets/pc4mb3h8hz/1

[D38] ODIR. ODIR Challenge, Intelligent Ophthalmology. https://odir2019.grand-challenge.org/dataset/

[D39] ORIGA650. ORIGA(-light): an online retinal fundus image database for glaucoma analysis and research. https://doi.org/10.6084/m9.figshare.24549217

[D40] PALM1200. PALM: PAthoLogic Myopia Challenge. https://ieee-dataport.org/documents/palm-pathologic-myopia-challenge

[D41] PAPILA. PAPILA: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. https://figshare.com/articles/dataset/PAPILA/14798004/1?file=28454352

[D42] PARAGUAY. Dataset from fundus images for the study of diabetic retinopathy. https://zenodo.org/records/4647952

[D43] Papilledema. Machine Learning For Pseudopapilledema. https://doi.org/10.17605/OSF.IO/2W5CE

[D44] REFUGE1200. REFUGE Challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. https://ieee-dataport.org/documents/refuge-retinal-fundus-glaucoma-challenge

[D45] RFMiD2. Retinal Fundus Multi-Disease Image Dataset (RFMiD) 2.0: A Dataset of Frequently and Rarely Identified Diseases. https://zenodo.org/records/7505822

[D46] RFMiD3200. Retinal Fundus Multi-Disease Image Dataset (RFMiD): A Dataset for Multi-Disease Detection Research. https://riadd.grand-challenge.org/download-all-classes/

[D47] RIM-ONE. RIM-ONE DL: A Unified Retinal Image Database for Assessing Glaucoma Using Deep Learning. https://github.com/miag-ull/rim-one-dl

[D48] RITE40. Automated Separation of Binary Overlapping Trees in Low-Contrast Color Retinal Images. https://eye.medicine.uiowa.edu/rite-dataset

[D49] ROD. Retinal Occlusion Dataset. https://www.kaggle.com/datasets/gracemariabinu/retinal-occlusion-dataset

[D50] ROI. Accuracy Assessment of Intra- and Intervisit Fundus Image Registration for Diabetic Retinopathy Screening. https://www.re3data.org/repository/r3d100012445

[D51] STARE. STARE: Structured Analysis of the Retina. https://cecas.clemson.edu/~ahoover/stare/

[D52] SUSTech-SYSU. The sustech-sysu dataset for automated exudate detection and diabetic retinopathy grading. https://figshare.com/s/792d1b02f65be0c08214

[D53] ToxoFundus. Dataset of fundus images for the diagnosis of ocular toxoplasmosis. https://zenodo.org/records/13883485

[D54] VietAI. VietAI Advanced Course—Retinal Disease Detection. https://www.kaggle.com/competitions/vietai-advance-retinal-disease-detection-2020/data