

Enhancing diagnostic accuracy in rare and common fundus diseases with a knowledge-rich vision-language model

Received: 27 September 2024

Accepted: 27 May 2025

Published online: 01 July 2025

 Check for updates

Meng Wang  ^{1,2,30}, Tian Lin  ^{3,30}, Aidi Lin  ³, Kai Yu ⁴, Yuanyuan Peng ⁵, Lianyu Wang ⁶, Cheng Chen ⁷, Ke Zou ⁸, Huiyu Liang ³, Man Chen ³, Xue Yao ³, Meiqin Zhang ³, Binwei Huang ³, Chaoxin Zheng ³, Peixin Zhang ³, Wei Chen ³, Yilong Luo ³, Yifan Chen ³, Honghe Xia ³, Tingkun Shi ³, Qi Zhang ³, Jinming Guo ³, Xiaolin Chen ³, Jingcheng Wang ⁹, Yih Chung Tham  ^{1,2}, Dianbo Liu ^{1,2}, Wendy Wong ^{1,2}, Sahil Thakur  ¹⁰, Beau J. Fenner  ^{10,11}, Danqi Fang ¹², Siying Liu ¹³, Qingyun Liu ¹³, Yuqiang Huang ³, Hongqiang Zeng ¹⁴, Yanda Meng ¹⁵, Yukun Zhou  ^{16,17,18}, Zehua Jiang ^{19,20}, Minghui Qiu ²¹, Changqing Zhang ²², Xinjian Chen  ^{23,24}, Sophia Y. Wang  ²⁵, Cecilia S. Lee  ^{26,27}, Lucia Sobrin ²⁸, Carol Y. Cheung  ¹², Chi Pui Pang ^{3,12}, Pearse A. Keane  ^{17,18}, Ching-Yu Cheng  ^{1,2,10,11}  ³  & Huazhu Fu  ²⁹ 

Previous foundation models for fundus images were pre-trained with limited disease categories and knowledge base. Here we introduce RetiZero, a vision-language model that incorporates knowledge from over 400 fundus diseases. The model is pre-trained on 341,896 fundus images with accompanying text descriptions gathered from diverse sources across multiple ethnicities and countries. RetiZero demonstrates exceptional performance across various downstream tasks including zero-shot disease recognition, image-to-image retrieval, clinical diagnosis assistance, few-shot fine-tuning, and cross-domain disease identification. In zero-shot scenarios, it achieves Top-5 accuracies of 0.843 for 15 diseases and 0.756 for 52 diseases, while for image-to-image retrieval, it scores 0.950 and 0.886 respectively. Notably, RetiZero's Top-3 zero-shot performance exceeds the average diagnostic accuracy of 19 ophthalmologists from Singapore, China, and the United States. The model particularly enhances clinicians' ability to diagnose rare fundus conditions, highlighting its potential value for integration into clinical settings where diverse eye diseases are encountered.

Blindness and visual impairment represent a substantial disease burden globally, impacting millions of individuals across all populations. Detection and timely treatment of ocular conditions, such as retinal and optic nerve diseases, are crucial for reducing severe and permanent damage. However, the insufficient availability of ophthalmic medical resources severely limits the prompt screening and

management of fundus diseases with vast regional differences in many parts of the world.

In recent years, artificial intelligence (AI)-based fundus disease screening systems have been proposed and achieved promising performance on fundus disease detection and patients' referral. Nevertheless, most previous AI-based methods were customized for and

limited to specific diseases, such as diabetic retinopathy (DR)^{1,2}, glaucoma^{3,4}, and retinopathy of prematurity^{5,6}. Although several methods were proposed for simultaneously screening multiple fundus diseases with promising performance⁷⁻⁹, most current AI models for ocular disease screening were trained on task-specific datasets, leading to inevitable errors in detection when there were new data (e.g., images acquired by different camera) or changes in tasks (e.g., introducing new or rare categories). Furthermore, due to limited healthcare resources and the varying prevalence of fundus disease, collecting comprehensive datasets covering all kinds of fundus abnormalities is time-consuming and challenging. Consequently, most AI models were trained on limited data and disease categories, restricting their feature representation. Applying these models to different real-world settings or tasks requires extensive retraining with large datasets. Moreover, data quality and labeling issues further limit the widespread adoption of AI models in ophthalmic clinical settings, especially from a global perspective.

Driven by the abundance of big data and robust computing hardware, large foundation models (LFMs) have excelled in computer vision tasks^{10,11}. Pre-trained on massive datasets, LFs provide rich feature support for downstream tasks, such as object detection¹², few-shot recognition¹³, and zero-shot¹⁴, etc. The first ophthalmic LFM, RETFound¹⁵, introduced in 2023, was trained on large, unannotated retinal images using a masked autoencoder (MAE) framework¹⁶. It provides rich feature support and improves the performance of downstream tasks. However, such an approach can hinder the model's capacity to align feature information with labels in downstream tasks. In contrast, the Foundation LAnguage-Image model of the Retina (FLAIR)¹⁷, a Contrastive Language-Image Pre-training (CLIP)-based LFM, enhance feature representation by aligning text descriptions with image features, improving feature-label alignment but having difficulties with complex semantic features in medical imaging¹⁰. MAE-based pretraining approaches excel in capturing complex semantic features in medical imaging by leveraging masked autoencoding techniques that focus on reconstructing obscured regions of an image, thereby fostering a deep understanding of local structures and subtle pathological details essential for accurate diagnostics. This approach encourages the model to learn rich, fine-grained representations by emphasizing contextual and structural information within the image. In contrast, CLIP-based pretraining primarily optimizes for global image-text alignment, aligning entire images with their corresponding textual descriptions without delving into the intricate internal features. While CLIP models are effective for tasks requiring broad semantic understanding and cross-modal associations, their emphasis on image-level alignment limits their ability to discern and interpret the nuanced and complex semantic patterns crucial in medical contexts. Consequently, CLIP-based methods struggle to effectively handle the detailed and sophisticated features necessary for precise medical image analysis, suggesting a significant limitation in their use for healthcare diagnostics. Furthermore, current LFs for ophthalmic imaging are pre-trained on extensive yet categorically limited datasets. Therefore, developing LFs with comprehensive ophthalmic disease knowledge would be crucial for representing complex retinal features to enhance downstream task performance. Nevertheless, collecting massive and diverse ophthalmic data that covers a wide range of fundus diseases for pretraining remains a significant challenge.

To address these problems and challenges, we collected 341,896 fundus images-text pairs from 29 publicly available datasets (containing 303,124 fundus images with labels), 180 ophthalmic literature (23,328 fundus images with diseases-related keywords), and online resources (15,544 fundus image-text pairs), encompassing over 400 retinal and optic nerve diseases across multiple countries, regions and ethnicities (Fig. 1a). As shown in Fig. 1b, our LFM, RetiZero, is based on

a contrastive vision-language pretraining framework that integrates MAE-based pretraining knowledge and low-rank training methods. Moreover, we introduced an uncertainty vision-language feature calibration method using Dirichlet reparameterization within the contrastive vision-language pretraining framework, to further align vision and language features in the high-dimensional embedding space. Consequently, RetiZero achieved superior performance in various downstream tasks, including zero-shot fundus disease recognition, image-to-image fundus disease retrieval, AI-assisted clinical diagnosis, internal domain fundus disease identification, few-shot fine-tuning, and cross-domain fundus disease identification.

Results

Zero-shot fundus disease recognition

The biggest advantage of RetiZero is the capability of zero-shot learning, which enables RetiZero to recognize fundus diseases using only textual prompts, without needing to retrain or fine-tune the model with labeled fundus images (Task I in Fig. 1c). As shown in Fig. 2a, RetiZero achieved overall Top-1, Top-3, and Top-5 scores of 0.442, 0.702, 0.840, respectively, for recognizing 15 common fundus diseases and normal condition of 30,089 fundus images (Eye-15 dataset, see in Supplementary Table 1). These scores of RetiZero improved by 25.5% for Top-1 ($P=0.01$), 15.7% for Top-3 ($P=0.03$), and 15.6% for Top-5 ($P=0.04$) over FLAIR. Furthermore, in the analysis of individual diseases, RetiZero showed remarkable zero-shot capability in identifying most categories, especially for retinitis pigmentosa (Top-1: 0.819, Top-3: 0.953, and Top-5: 0.974), retinal detachment (Top-1: 0.762, Top-3: 0.906, and Top-5: 0.963), and glaucoma (Top-1: 0.748, Top-3: 0.929, and Top-5: 0.972) (Supplementary Fig. 1). To further validate RetiZero's zero-shot capability in more challenging clinical scenarios, we assembled a more demanding dataset named EYE-52 (see in Supplementary Table 2). This dataset included 7007 fundus images from various ophthalmology clinics, covering 52 fundus diseases. Many of these diseases are rare in eye clinics but can lead to severe visual impairment if left undiagnosed. The incidence and prevalence of each category in the EYE-52 dataset were shown in Supplementary Table 3. As depicted in Fig. 2b, RetiZero achieved overall Top-1, Top-3, and Top-5 scores of 0.360, 0.626, and 0.756, respectively, for recognizing these 52 types of fundus diseases in a zero-shot manner, providing superior performance compared FLAIR (0.092, 0.263, and 0.340, respectively) and Random recognizing (0.029, 0.088, and 0.147, respectively). Furthermore, RetiZero demonstrated superior zero-shot performance, especially for recognizing several rare fundus diseases. For instance, RetiZero achieved Top-1, Top-3, and Top-5 scores of 0.616, 0.791, and 0.861, respectively, for identifying Bietti Crystalline dystrophy; and 0.509, 0.808, and 0.915, respectively for recognizing chorioretinal coloboma, the Top-1, Top-3, and Top-5 scores were (Supplementary Fig. 2). Figure 2c shows the Top-5 prediction results provided by RetiZero and FLAIR for three rare fundus disease samples. More details on the rest of the 52 disease categories can be found in Supplementary Fig. 2.

Fundus disease identification by image-to-image retrieval

For image-to-image retrieval (Fig. 1d, Task II), we treated each fundus image in turn as the query, removing it from the main set to form the candidate pool of all remaining images. We then passed the query through RetiZero's image encoder, generating a feature embedding. Using the same encoder, we extracted embeddings for every image in the candidate pool and computed similarity scores between the query's embedding and each candidate's embedding. These similarity scores ranked the candidate images by how closely they matched the query, allowing us to retrieve the Top-K matches (e.g., Top-5). Figure 2d and Supplementary Fig. 3 illustrate the excellent performance of RetiZero in identifying 15 fundus diseases through image-to-image retrieval. The overall scores for Top-1, Top-3, and Top-5 are 0.854, 0.928, and 0.950, respectively, representing an improvement of

9.4%, 4.8%, and 3.2% over RETFound (all $P < 0.001$), and 300.2%, 121.1%, 74.0% over FLAIR (all $P < 0.001$). In addition, RetiZero demonstrated the best performance across all categories compared to RETFound and FLAIR (Supplementary Fig. 3). In the more challenging Eye-52 dataset, RetiZero achieved overall Top-1, Top-3, and Top-5 scores of 0.726, 0.843, and 0.886, respectively (Fig. 2e). These scores represented improvements of 12.4%, 7.9%, and 6.3% over RETFound (all $P < 0.001$), and 767.8%, 389.9%, 271.5% over FLAIR, respectively (all $P < 0.001$). Figure 2f shows an example of the Top-5 prediction results from RetiZero, RETFound, and FLAIR. Furthermore, in the analysis of individual diseases, RetiZero demonstrated remarkable potential, particularly in identifying several rare fundus diseases, such as punctate

inner choroidopathy multifocal choroiditis (Top-1: 0.902, Top-3: 0.946, and Top-5: 0.962), chorioretinal coloboma (Top-1: 0.819, Top-3: 0.893, and Top-5: 0.910), and Bietti crystalline dystrophy (Top-1: 0.861, Top-3: 0.936, and Top-5: 0.942). More details on the 52 disease categories can be found in Supplementary Fig. 4. In addition, we calculated Precision@1, Precision@3, and Precision@5 to comprehensively evaluate RetiZero's performance in the task of fundus disease identification through image-to-image retrieval. Figure 2d, e demonstrates that RetiZero achieved the highest Precision@1, Precision@3, and Precision@5 on both testing datasets, EYE-15 and EYE-52. Meanwhile, RetiZero demonstrated the best performance across most categories compared to RETFound and FLAIR (Supplementary Figs. 5 and 6).

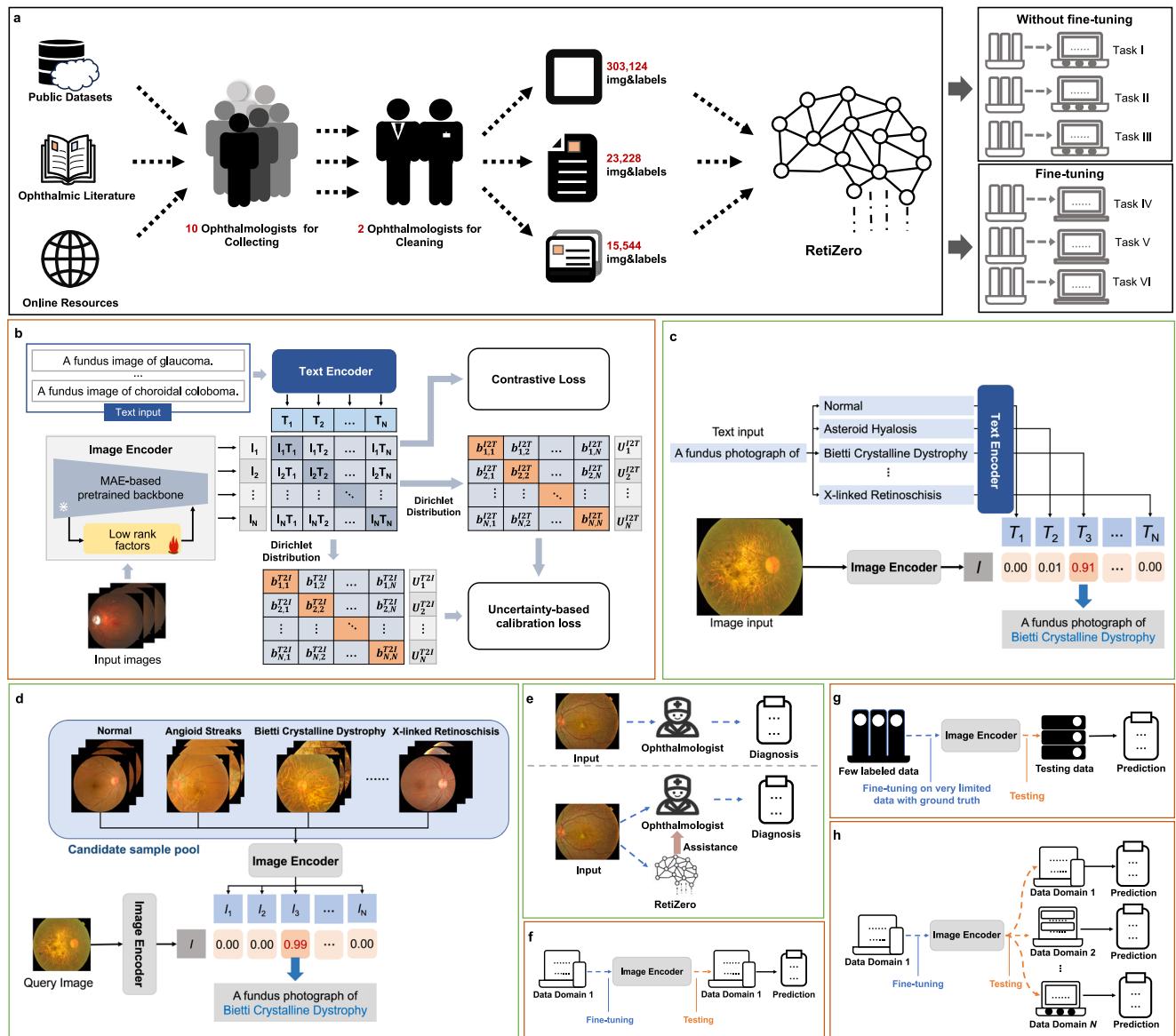


Fig. 1 | Overview of the framework. **a** Datasets for RetiZero pretraining: The RetiZero model was pre-trained using data from three primary sources: public datasets, ophthalmic literature, and online resources. We assembled a team of 12 ophthalmologists for manual data collection and cleaning. This involved downloading images and corresponding labels from public datasets, extracting images and corresponding disease-related keywords from ophthalmic literature, and downloading retinal diseases-relevant image-text pairs from online resources. **b** RetiZero, which combines the strengths of self-supervised learning based on the MAE architecture and contrastive learning from the CLIP architecture. Moreover, we introduce an uncertainty vision-language feature calibration method into the

contrastive vision-language pretraining framework, to further calibrate visual-language features in the high-dimensional embedding space. **c** Task I: Zero-shot fundus disease recognition. **d** Task II: Fundus disease identification by image-to-image retrieval. **e** Task III: AI-assisted clinical diagnosis. **f** Task IV: Internal domain retinal disease identification. “Internal domain” means that we fine-tuned and tested the model using the data with similar feature distribution. **g** Task V: Few-shot fine-tuning. We evaluate RetiZero’s performance in identifying fundus diseases with very limited training data. **h** Task VI: Cross-domain fundus disease identification. “Cross-domain” means that we fine-tuned and tested the model using the data with different feature distributions.

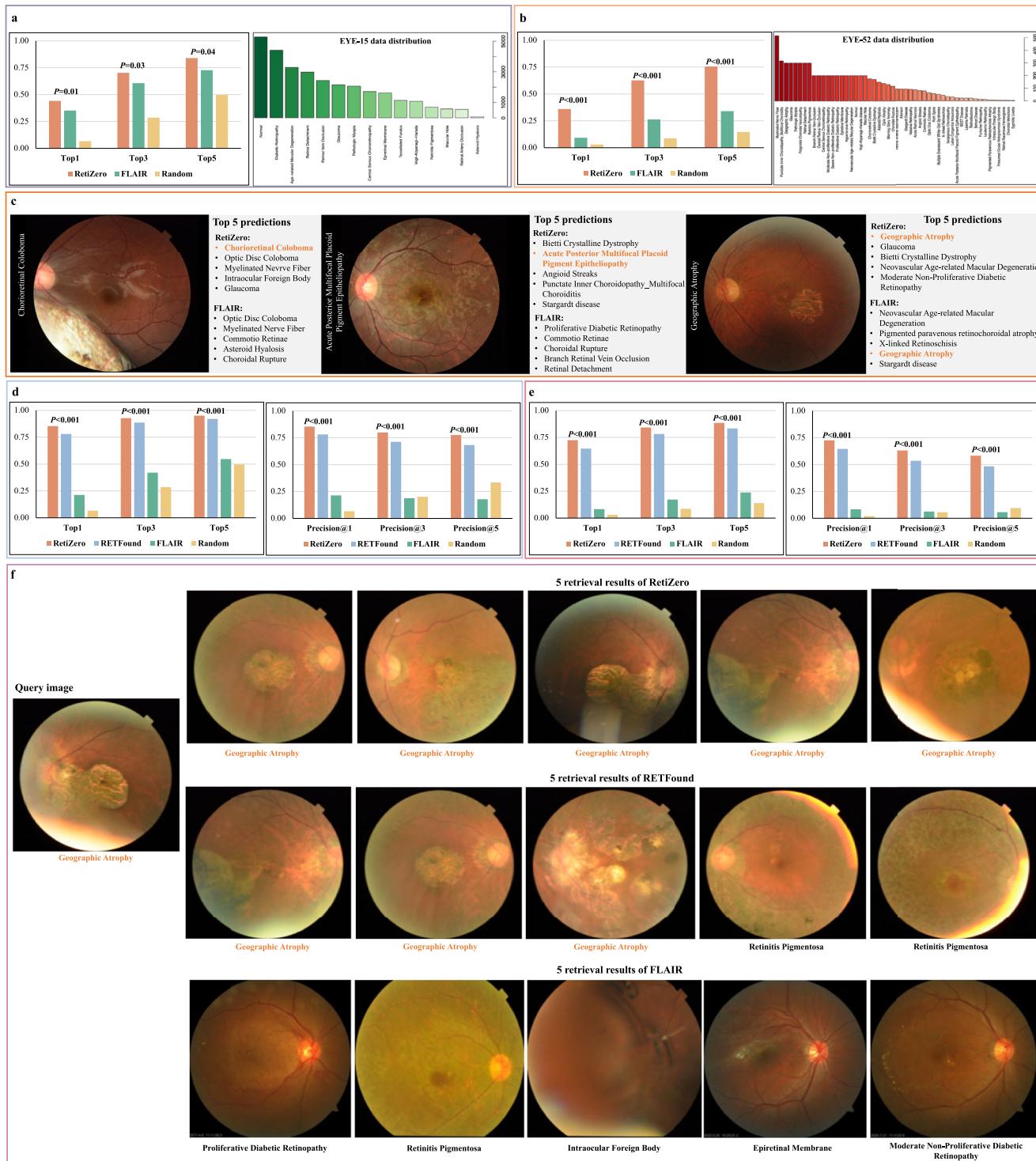


Fig. 2 | Overall Top-1, Top-3, and Top-5 scores for zero-shot based fundus disease recognition and Fundus disease identification by image-to-image retrieval. **a** The zero-shot performance on EYE-15 dataset, which contains 30,089 fundus images including 14 common fundus diseases and a normal condition. **b** The zero-shot performance on the EYE-52 dataset, which contains 7007 fundus images

including 51 categories of fundus diseases and a normal condition. **c** Zero-shot fundus diseases identification samples. **d** Image-to-image retrieval performance on EYE-15 dataset. **e** Image-to-image retrieval performance on the EYE-52 dataset. **f** Image-to-image retrieval samples. All P values were calculated with the two-sided t-test with 95% confidence intervals. Source data are provided as a Source Data file.

Furthermore, we further visualized the heatmaps of different models' weights for various fundus diseases using GradCAM¹⁸. Supplementary Fig. 7 presents heatmaps illustrating the weights of different foundational models for various fundus diseases. RetiZero's weights were more precisely concentrated on the regions affected by different fundus diseases.

AI-assisted clinical diagnosis

As illustrated in Task III in Fig. 1e, we first compared the performance of RetiZero with that of 19 ophthalmologists from Singapore, China and the U.S., and then examined whether RetiZero can assist in the clinical diagnosis made by ophthalmologists. In brief, for this clinical validation study, we created a subset of data comprising a total of

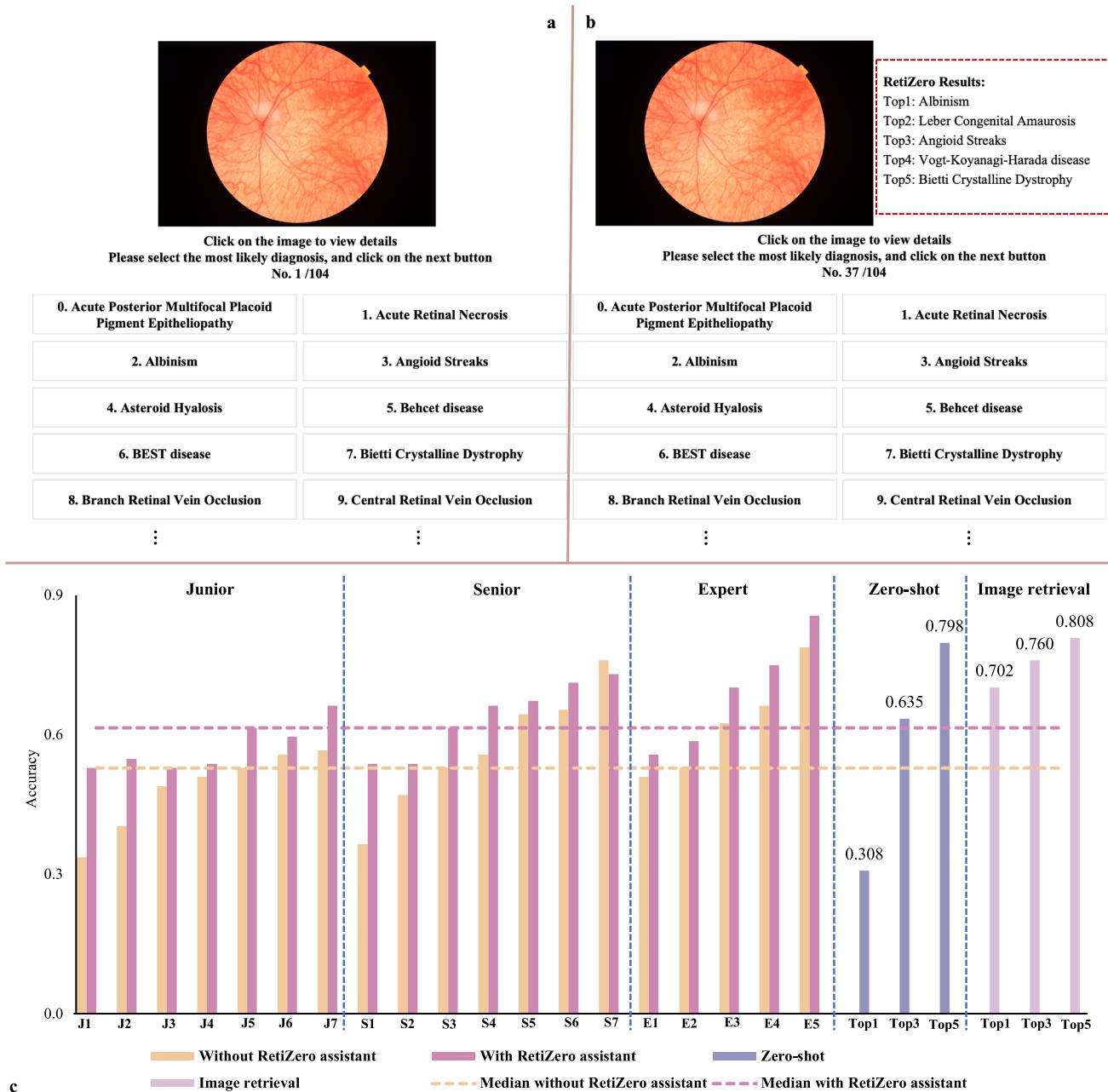


Fig. 3 | AI-assisted clinical diagnosis results. a Online fundus image reading system without RetiZero assistance. **b** Online fundus image reading system with RetiZero assistance. **c** Ophthalmologist diagnostic results, Top-1, Top-3, and Top-5

performance for zero-shot and image-to-image retrieval. Source data are provided as a Source Data file.

104 images by randomly selecting 2 samples per category from the EYE-52 set, and the ophthalmologists were asked to make a clinical diagnosis based on the retinal photographs without (Round 1, Fig. 3a) and with (Round 2, Fig. 3b, conducted one week after Round 1) the assistance of RetiZero. The diagnostic accuracy of the 19 ophthalmologists ranged from 0.337 to 0.788, with the median of 0.582, while RetiZero's zero-shot Top-1, Top-3, and Top-5 accuracies are 0.308, 0.635, and 0.798, respectively (Fig. 3c). Therefore, RetiZero's zero-shot Top-3's performance was comparable to that of the majority of ophthalmologists, while its Top-5's performance surpassed that of all ophthalmologists. Furthermore, RetiZero's performance the fundus disease identification by image-to-image retrieval for RetiZero achieved Top-1, Top-3, and Top-5 accuracies of 0.702, 0.760, and 0.808, surpassing that of the majority of ophthalmologists.

With the assistance of RetiZero, the performance of 18 out of the 19 ophthalmologists improved (Fig. 3c), out of 1976 total responses (104 questions \times 19 doctors), 1569 responses (79.4%) remained unchanged, while 407 responses (20.6%) were modified in the second round. Of the 407 modified responses, 279 (68.6%) were changed from incorrect to correct, demonstrating that most AI-assisted modifications helped the clinicians rectify previous misdiagnoses. Moreover, as shown in Table 1, we further grouped the ophthalmologists by years of experience as junior (\leq 5 years, 7 doctors), senior (5–10 years, 7 doctors), and expert ($>$ 10 years, 5 doctors). Their average first-round accuracies were 48.5%, 56.9%, and 62.3%, respectively, improving to 57.4%, 63.9%, and 69.0% in the second round, with respective improvements of 18.4% ($P = 0.01$), 12.3% ($P = 0.03$), and 10.8% ($P < 0.001$). This demonstrates that RetiZero-assisted diagnosis enhances accuracy across all experience levels, with junior clinicians

Table 1 | Details for clinical evaluation

Ophthalmologist	Junior (Confidence level)		Senior (Confidence level)		Expert (Confidence level)	
	Without Reti-Zero assistance	With RetiZero assistance	Without Reti-Zero assistance	With RetiZero assistance	Without Reti-Zero assistance	With RetiZero assistance
1	0.337 (Low)	0.529 ^a (Low)	0.365 (Very Low)	0.538 ^a (Low ^a)	0.510 (Low)	0.558 ^a (Low)
2	0.404 (Low)	0.548 ^a (Low)	0.471 (High)	0.538 ^a (High)	0.529 (High)	0.587 ^a (High)
3	0.490 (Very Low)	0.529 ^a (Very Low)	0.529 (Moderate)	0.615 ^a (Moderate)	0.625 (Moderate)	0.702 ^a (Moderate)
4	0.510 (Moderate)	0.538 ^a (Moderate)	0.558 (Moderate)	0.663 ^a (High ^a)	0.663 (High)	0.750 ^a (High)
5	0.529 (Very Low)	0.615 ^a (Very Low)	0.644 (Moderate)	0.673 ^a (High ^a)	0.788 (Moderate)	0.856 ^a (Moderate)
6	0.558 (Moderate)	0.596 ^a (Moderate)	0.654 (High)	0.712 ^a (High)	/	/
7	0.567 (Moderate)	0.663 ^a (High ^a)	0.760 (Moderate)	0.731 ^b (Moderate)	/	/
Mean	0.485	0.574 ^a	0.569	0.639 ^a	0.623	0.691 ^a
Improvement	18.4%	12.3%				10.8%
Overall mean		Without RetiZero assistance		With RetiZero assistance		
	0.552			0.628		
Overall mean improvement				13.8%		
Statistics of answering modification						
RetiZero	Doctors	False→True (Proportion of modified samples)	True→False (Proportion of modified samples)		Total (Proportion of modified samples)	
True		256 (62.9%)	89 (21.9%)		345 (84.8%)	
False		23 (5.7%)	39 (9.6%)		62 (15.2%)	
Total		279 (68.6%)	128 (31.4%)		407 (100%)	

^aIndicates an improvement, while ^bindicates a decline. Source data are provided as a Source Data file.

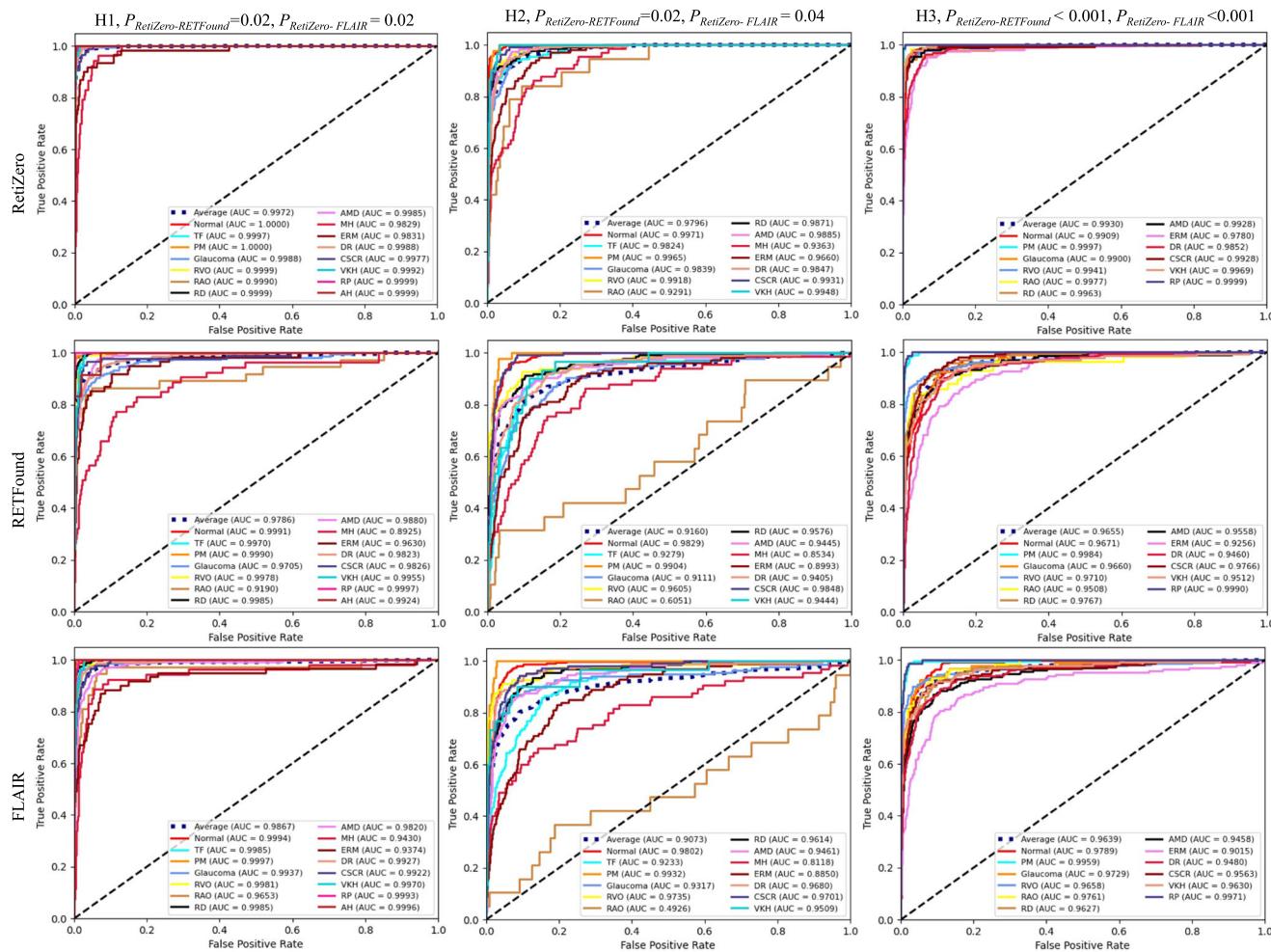


Fig. 4 | The receiver operating characteristic (ROC) curves for internal domain retinal disease identification. The P values were calculated with the two-sided t-test, with 95% confidence intervals. AH Asteroid Hyalosis, AMD Age-related Macular Degeneration, CSCR, Central Serous Chorioretinopathy, DR Diabetic Retinopathy,

ERM Epiretinal Membrane, MH Macular Hole, PM Pathologic Myopia, RAO Retinal Artery Occlusion, RD Retinal Detachment, RP Retinitis Pigmentosa, RVO Retinal Vein Occlusion, TF Tessellated Fundus, VKH Vogt-Koyanagi-Harada disease. Source data are provided as a Source Data file.

benefiting the most. Additionally, the clinicians' confidence in their diagnoses increased from an average of 2.7 (low to moderate) in the first round to 3.0 (moderate, $P=0.04$) in the second round, suggesting that RetiZero assistance not only improved accuracy but also boosted their confidence in making clinical diagnoses. Finally, there was a strong correlation between the top-ranking scores from RetiZero and the response modification scores of ophthalmologists ($r=0.614$, $P<0.001$, see Methods for details), suggesting that the higher the correct result was ranked within the model's top 5 predictions, the greater the probability that the doctor would arrive at an accurate diagnosis. Moreover, to more clearly illustrate the effectiveness of RetiZero in assisted diagnostics, we further provided additional quantitative analysis in Supplementary Figs. 8–10.

Internal domain fundus disease identification

In Task IV (Fig. 1f), we collected three clinical datasets of retinal photographs, named as H1, H2, and H3, to validate the performance of RetiZero in internal domain fundus disease identification tasks. Supplementary Fig. 11 provides the data collection process and annotation details of three datasets. “Internal domain” means that we fine-tuned and validated the model separately within each of the three datasets. The details of three datasets are shown in Supplementary Tables 4–6. As shown in Fig. 4, RetiZero achieved average AUCs of 0.997, 0.980, and 0.993 on the three datasets, respectively, each encompassing 15, 13, and

12 different categories of fundus diseases and/or normal condition, respectively. These results represent improvements of 1.9% ($P=0.02$), 6.9% ($P=0.02$) and 2.9% ($P<0.001$) compared to RETFound, and 1.1% ($P=0.02$), 8.0% ($P=0.04$) and 3.0% ($P<0.001$) compared to FLAIR. This is particularly apparent for certain fundus diseases with specific features, such as macular hole, epiretinal membrane, and retinal artery occlusion.

Few-shot fine-tuning

In task V (Fig. 1g), we fine-tuned the model using only five images from each fundus disease to evaluate RetiZero's performance with very limited training data. The data details are provided in Supplementary Tables 7–9. RetiZero achieved the highest AUC scores across the three datasets compared to RETFound and FLAIR (Fig. 5). In the task of identifying 15, 13, and 12 types of fundus diseases in the H1, H2, and H3 dataset, RetiZero achieved AUC values of 0.967, 0.859, and 0.942 respectively, representing improvements of 7.2% to 35.1% over RETFound and FLAIR (all $P<0.001$). These results indicated that even with limited annotated data samples, RetiZero can effectively learn the characteristic information of different fundus diseases in fundus images.

Cross-domain fundus disease identification

To assess the robustness of RetiZero in the task of cross-domain fundus disease identification (Task VI, Fig. 1h), we reorganized the three

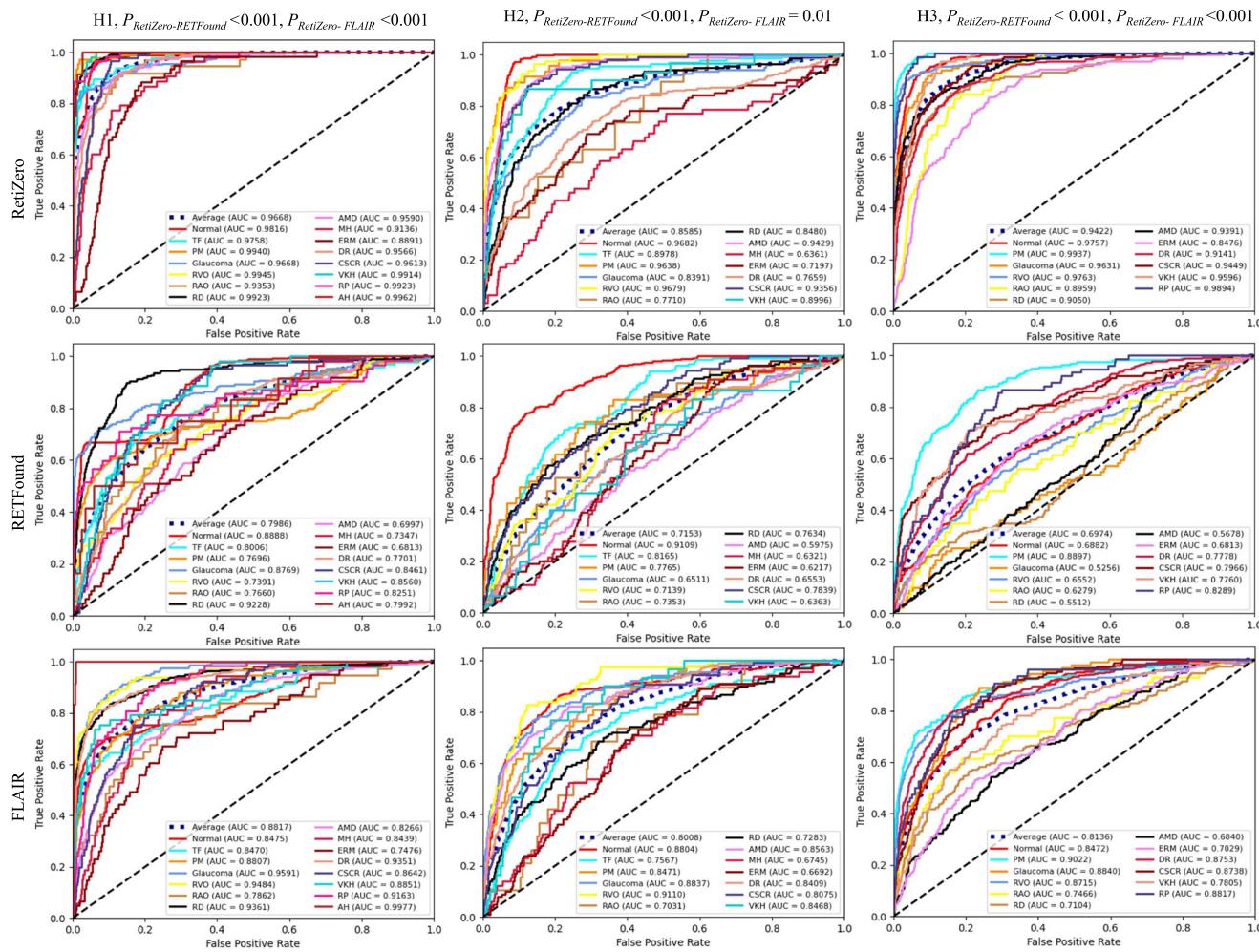


Fig. 5 | The receiver operating characteristic (ROC) curves for few-shot learning. The P values were calculated with the two-sided t-test, with 95% confidence intervals. AH Asteroid Hyalosis, AMD Age-related Macular Degeneration, CSCR Central Serous Chorioretinopathy, DR Diabetic Retinopathy, ERM Epiretinal

Membrane, MH Macular Hole, PM Pathologic Myopia, RAO Retinal Artery Occlusion, RD Retinal Detachment, RP Retinitis Pigmentosa, RVO Retinal Vein Occlusion, TF Tessellated Fundus, VKH Vogt-Koyanagi-Harada disease. Source data are provided as a Source Data file.

datasets of H1, H2, and H3 and only used the data with shared diseases categories across these datasets. Then we sequentially used one of the reorganized datasets, named rH1, rH2, and rH3, as internal testing sets and utilized the remaining two datasets as external testing sets to verify the robustness of different foundation models. The data information for different experimental strategies is presented in Supplementary Tables 10–12. As shown in Fig. 6, RetiZero achieved promising performance in all validation settings. In the internal testing set of the three datasets, RetiZero achieved AUC values of 0.998, 0.986, and 0.990, respectively, representing significant improvements of 3.1%, 8.6%, and 6.3% over RETFound (all $P < 0.001$); and of 0.7%, 4.4%, and 5.0% over FLAIR (all $P < 0.001$, Fig. 6a). In the external testing sets, the performance of RetiZero was similar to the internal testing set, with all AUCs ≥ 0.912 , and significantly outperformed RETFound and FLAIR in all tasks (all $P \leq 0.01$, see in Fig. 6b, c). Additionally, RetiZero achieved outstanding performance in the identification of fundus diseases across most of the categories, especially in diseases with unique pathologic features such as epiretinal membrane, retinal artery occlusion, and central serous chorioretinopathy (Supplementary Figs. 12–14). We further conducted additional experiments to assess whether training with both domains rH1 and rH2, as opposed to training with only one of them, yields improved performance on domain rH3 (see Supplementary Fig. 15). The results indicate that incorporating both rH1 and rH2 during training leads to performance

enhancements of 0.5% and 0.6% on rH3, with statistical significance ($P = 0.029$ and $P = 0.008$, respectively). These findings highlight the advantages of domain combination in achieving better generalization. Furthermore, we evaluated the inference times of various foundational models for diagnosing a single image (refer to Supplementary Table 13). RetiZero, RETFound, and FLAIR demonstrated inference times of 0.013, 0.012, and 0.017 seconds per image, respectively. Although RetiZero's inference time is marginally higher than RETFound's, it remains well within the acceptable range for real-time processing.

In addition, we rigorously tested our model's performance across multiple datasets representing diverse ethnic groups. This included two publicly available datasets (See Supplementary Tables 14, 15): the 2019 Sydney Innovation Challenge dataset (SIC, encompassing Caucasian and Indian populations, etc.) and the SUSTech dataset (was collected from China) for Diabetic Retinopathy (DR) staging. Additionally, we utilized two in-house datasets for AMD and DR screening (See Supplementary Tables 16, 17): the Singapore Malay Eye Study (SiMES) dataset was collected from Malay individuals and the Singapore Indian Eye Study (SINDI) dataset was collected from Indian individuals. Across these diverse populations, RetiZero consistently outperformed RETFound and FLAIR (see Supplementary Fig. 16). These comprehensive experimental results underscore RetiZero's robust performance and generalizability, affirming its efficacy in

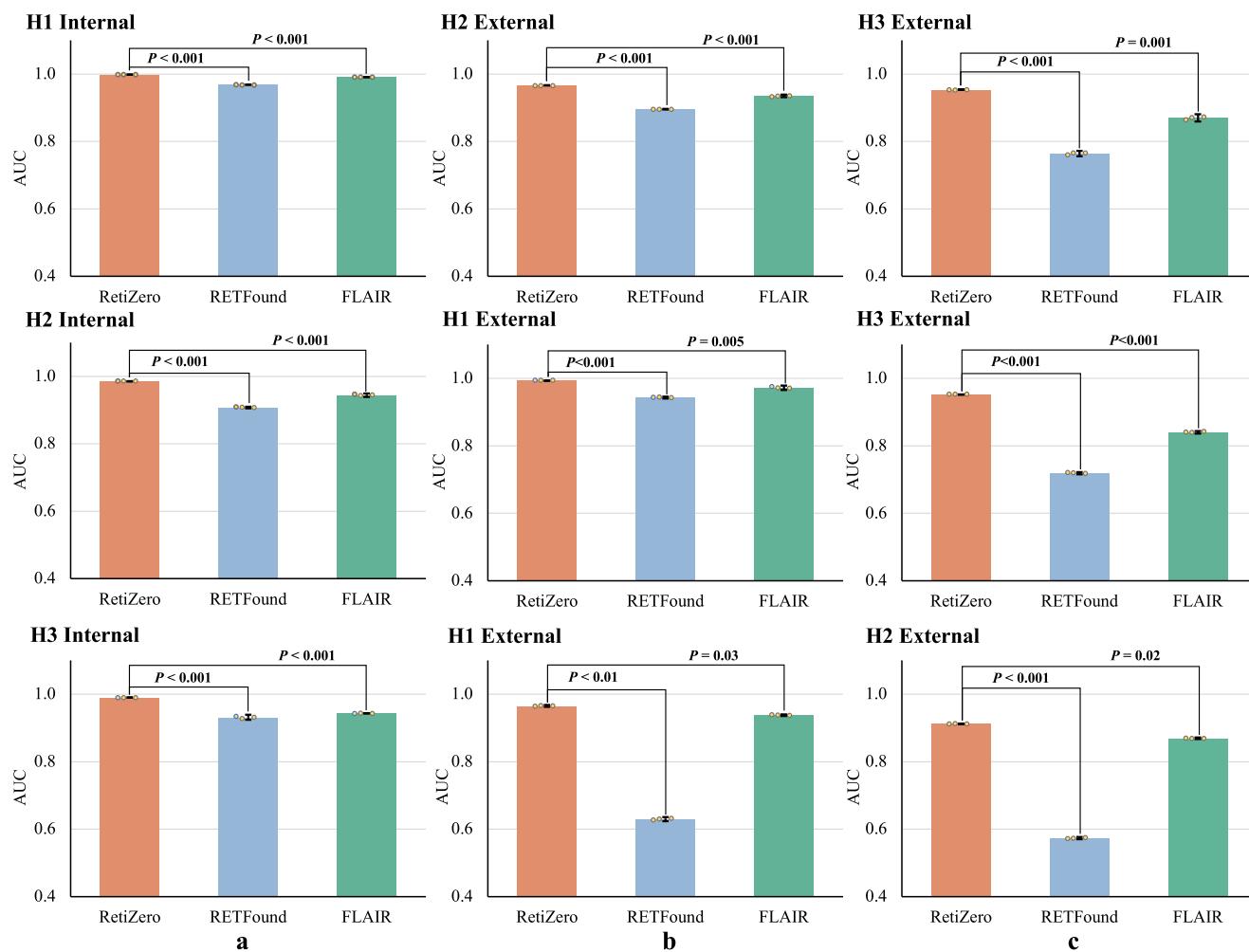


Fig. 6 | Cross-domain performance (AUC) of different foundation models for fundus disease screening. Column **a**, Internal evaluation: Different foundation models were adapted to each dataset by fine-tuning and internally evaluated on hold-out testing data. Columns **b**, **c**, performance on external validation sets: The three foundation models were tested on the other two external validation datasets. *P* value was calculated using the two-sided t-test with 95% confidence intervals. The

dots represent the AUC scores from fine-tuning experiments with three different random seeds. The error bars show 95% confidence intervals, and the bar center represents the mean value of the AUC. The disease categories and dataset strategy information are listed in Supplementary Tables 10–12. Source data are provided as a Source Data file.

accurately diagnosing fundus diseases across varied demographic groups and enhancing its potential for widespread clinical application.

Discussion

In this study, we trained a vision-language-foundation model, RetiZero, using a vast amount of image-text pairs. Comprehensive experimental results demonstrated that RetiZero has strong capability in representing fundus disease features across a wide range of downstream tasks of fundus disease identification, including zero-shot recognition, image-to-image retrieval, AI-assisted clinical diagnosis, internal domain and cross domain classification, and few-shot fine-tuning. The performance of RetiZero is superior to two state-of-the-art ophthalmic LMs, RETFound¹⁵ and FLAIR¹⁷. These results collectively demonstrated the superior generalizable and robust performance of RetiZero in both common and rare fundus disease identification.

The superiority of RetiZero over RETFound¹⁵ and FLAIR¹⁷ can be attributed to its unique design and diverse data used for pre-training. Specifically, although the RETFound model¹⁵, pre-trained on a large number of fundus images using the MAE architecture, can enhance the performance of various downstream tasks, it includes limited number of fundus disease categories, particularly rare fundus diseases. In addition, it lacks the incorporation of textual information, resulting in

inadequate characterization of image feature attributes. This makes it unsuitable for text prompt-based zero-shot fundus disease detection tasks, thereby limiting its application in clinical practice, particularly for the identification of rare fundus diseases. In contrast, FLAIR¹⁷, based on the CLIP architecture, incorporates textual description information during network training to enhance the representation of image feature attributes. However, it was pre-trained on a very limited dataset of fundus disease knowledge, leading to poor performance in zero-shot recognition tasks for rare fundus diseases. Furthermore, FLAIR¹⁷ and other standard CLIP-based models^{10,11} lack guidance for learning information such as lesion contours and topological structures in images, resulting in low performance in fundus disease identification through image-to-image retrieval. In contrast to these existing foundation modes, RetiZero leverages the synergy between MAE-based self-supervision and contrastive alignment with a broad set of disease descriptions to enhance its disease recognition capabilities. By combining the detailed feature extraction of MAE, which excels in capturing visual details through unsupervised learning from large-scale image datasets, with the semantic richness of contrastive text-image alignment, RetiZero gains a stronger implicit representation of lesions. This approach not only retains the granularity of local visual features critical for identifying subtle disease markers but also aligns

these features with a diverse range of textual disease descriptors. Consequently, RetiZero achieves superior performance by effectively bridging the gap between high-level semantic understanding and low-level visual details, making it adept at recognizing a wide array of fundus conditions, both common and rare.

Developing artificial intelligence algorithms to assist in the clinical diagnosis of rare fundus diseases has been challenging. Although image classification with few-shot learning has shown promising results^{19,20}, it is challenging to collect enough samples for training for extremely rare diseases, especially considering there are more than 400 fundus diseases. Zero-shot learning would be a better approach for this scenario. In previous literatures, there are only few reports investigating zero-shot learning for fundus images, which only focus on diabetic retinopathy or image quality assessment, and struggle with rare conditions^{21,22}. The image-to-image retrieval approach is another possible solution. Current literature only investigated diabetic retinopathy, not rare diseases, and the results showed limited performance, possibly due to the lack of integration of textual information crucial for complex disease identification^{23–26}. To our knowledge, our study is the first one to use zero-shot learning and image-to-image retrieval to cope with the challenge of rare disease diagnosis. RetiZero leverages both the MAE and CLIP architectures to enhance feature representation from diverse datasets, achieving superior performance in zero-shot fundus disease identification and image-to-image retrieval tasks. By integrating the MAE and CLIP architectures, RetiZero bridges the semantic gap, incorporates knowledge of over 400 types of fundus diseases, and has been validated on a clinical dataset encompassing more than 50 types.

Diagnosis of rare diseases is also very challenging for clinicians. Our results show that the accuracy of junior, senior, and expert ophthalmologists was 0.485, 0.569, and 0.623, respectively, in classifying the testing set with 104 retinal photos of 52 categories. In a clinical scenario, clinicians may also list several diseases for differential diagnosis, and then integrate other information from history, ocular examination, and other investigations for differential diagnosis. Our study mimics this approach. Although the zero-shot Top-1 accuracy (0.360) of RetiZero was lower than that of the ophthalmologists, the zero-shot Top-3 (0.626) and Top-5 (0.756) accuracies of RetiZero were comparable to and exceeded those of most ophthalmologists, respectively. This approach serves as a valuable reference for ophthalmologist in differential diagnosis. Our results also showed that using the zero-shot Top-5 results from RetiZero significantly enhanced the ophthalmologists' diagnostic accuracy, with expert ophthalmologists improving by 10.8% and junior ophthalmologists by 18.4%. Moreover, the correlation between RetiZero's reference predictions and the accuracy improvements observed in ophthalmologists underscores its potential to assist in differential diagnoses. In the exploration of AI in the medical field, initial studies primarily focused on comparing the performance of AI algorithms to physicians (AI-physician comparison) to the performance of AI algorithms^{27–29}. However, in real-world implementation, AI algorithms can make errors, posing potential risks to patient outcomes. To mitigate the risks, therefore, AI-assisted diagnosis or AI-physician collaboration have been proposed and investigated. Various approaches have been reported, such as using AI algorithms to suggest the most likely diagnosis, which physicians then confirm³⁰. Additionally, AI models have been shown to improve the diagnostic performance of junior doctors³¹. However, these studies mostly focus on common diseases with their AI models typically provide only one most probable disease diagnosis. In contrast, our RetiZero offers the top five results using the Zero-shot approach, significantly enhancing the diagnostic accuracy of ophthalmologists with various levels of clinical experiences. Beyond its robust zero-shot recognition capabilities, RetiZero's alignment of structured medical text with fundus imagery opens new avenues for practical clinical integration and cross-institutional collaboration. For

example, busy ophthalmologists can rapidly screen rare pathologies—such as Bietti crystalline dystrophy or chorioretinal coloboma—with pre-training on large labeled sets, facilitating timely triage or referral in resource-limited clinics. Its image-to-image retrieval function helps confirm diagnoses by surfacing visually similar cases from vast, heterogeneous archives, which is particularly beneficial when local datasets are small or unbalanced. Moreover, RetiZero's top-5 differential diagnosis output can enhance the confidence and accuracy of both junior and expert clinicians, making it especially attractive for teleophthalmology networks or multicenter screening programs.

We also recognized limitations and areas for improvements in the current work. Although our datasets include knowledge of over 400 types of fundus diseases, the imbalance across different categories may limit RetiZero's performance in downstream tasks. To address this, we plan to further enrich the dataset with a more balanced representation of various fundus diseases, particularly rare ones. Potentially, addressing data imbalance between rare and common ophthalmic pathologies could involve advanced synthetic data generation and tailored augmentation pipelines aimed specifically at underrepresented classes. Generative Adversarial Networks (GANs) or diffusion models can create realistic fundus images that mirror rare disease appearances, thereby expanding the sample diversity and allowing the network to learn clinically important features that are otherwise scarce^{32,33}. Beyond straightforward geometric transformations (e.g., rotations, flips), lesion-centered augmentations—like elastic deformations or localized color perturbations—further enrich minor-class variance without losing critical pathological signatures³⁴. These methods systematically enhance the representation of difficult or rare pathologies, ultimately bolstering model robustness in the face of severe data imbalance. In addition, while RetiZero has shown promising performance across multiple tasks and datasets, specialized models optimized for specific tasks may outperform more generalized models. Therefore, we will continue to explore targeted improvements to enhance RetiZero's performance for specific applications.

In conclusion, the proposed feature-calibrated fundus vision-language foundation model, RetiZero, which incorporates knowledge of over 400 fundus diseases, effectively captures the rich contextual feature in fundus images and learn the alignment between retinal image features and disease-related textual information. RetiZero demonstrated superior performance in feature representation and generalizability across different fundus disease recognition tasks, tested at multiple eye clinics using different fundus cameras, under different degrees of domain drift, and with very limited training samples. Notably, the excellent performance of RetiZero's exceptional performance in zero-shot fundus disease identification and image-to-image retrieval-based is of significant value for screening fundus diseases in clinical practice, particularly rare fundus conditions. Furthermore, comprehensive AI-assisted clinical diagnosis results further confirm that RetiZero can enhance ophthalmologists' diagnostic accuracy and confidence, offering particular benefit to less experienced clinicians.

Methods

This study was approved by the Joint Shantou International Eye Center Institutional Review Board and adhered to the principles of the Declaration of Helsinki. All fundus images used in this study were made de-identified. In accordance with IRB regulations, if the data does not contain any identifiable patient information, informed consent is not required. As a result, this study has been granted approval to waive the need for informed consent.

Data for pretraining

RetiZero uniquely integrates an MAE-based backbone with a CLIP-style contrastive framework and uncertainty-based feature calibration to achieve robust image-text alignment across more than 400 fundus

disease categories. RETFound was pre-trained using a Masked Auto-encoder (MAE) approach on over 900,000 fundus images. This large-scale pretraining allows RETFound to develop a deep understanding of fundus image features. Therefore, we utilized RETFound as the backbone of our image encoder. This approach allows us to enhance RetiZero's ability to effectively handle complex fundus image data by building on a robust, pre-established feature extractor. We pre-trained RetiZero using a dataset we collected, comprising of 341,896 image-text pairs that cover over 400 fundus diseases. As shown in Fig. 1a, the image-text pretraining data mainly consists of three parts: publicly available dataset with category information, data from the ophthalmic literature with disease-related keywords descriptions, and image-text pairs from online resources (Supplementary Table 18). Specifically, we collected a total of 303,124 fundus images from 29 publicly available datasets, encompassing over 100 distinct categories of fundus diseases. To ensure consistency in the presentation of image descriptions, we converted the fundus images with classification labels into a uniform textual description format (Fig. 1b). In addition, we engaged 12 ophthalmologists to compile and curate a comprehensive dataset. They reviewed 180 ophthalmic literature sources and online resources, identifying and documenting 23,228 fundus images with corresponding disease-related textual descriptions from the literature, as well as 15,544 fundus images from online resources. A team of ten ophthalmologists was assigned to this task. Their responsibilities included browsing various sources to locate fundus photographs, capturing screenshots of these images, and recording the source (literature/website), page number (book/page), and extracting the fundus description that uniquely corresponded to each respective fundus image. Our selection process was rigorous, including only those textual descriptions that precisely matched their corresponding color fundus photographs. Any information unrelated to the image, such as titles, background details, or descriptions of other imaging modalities or outside the field of view, was deliberately excluded. Furthermore, these ten ophthalmologists also assigned a diagnostic label for each image simultaneously. The final dataset underwent a thorough curation and validation process by two experienced ophthalmologists. Their responsibilities included unifying the labels to reduce label noise and excluding non-standard images, such as blurry, over- or under-exposed, montage, local view, monochrome images, and those annotated with additional graphics, thereby reducing image noise. Finally, all curated disease and lesion labels were converted into the standardized format "a fundus image of [disease/lesion labels]" and input into the text encoder model.

As shown in Supplementary Table 19, these images cover 414 ophthalmic labels, encompassing nearly all known fundus diseases to date. Meanwhile, Supplementary Table 20 provides further details about 180 ophthalmic literature sources. Furthermore, we collected 28,800 fundus images with relevant descriptions from the online resources. The 12 ophthalmologists manually cleaned and organized 15,544 images along with their corresponding disease-related textual descriptions. We pre-trained RetiZero using the PyTorch framework on an Nvidia Geforce DGX A100 GPU (80 G), with a batch size of 128 and the Adam optimizer. The data collection process for RetiZero pretraining is illustrated in Fig. 1a.

The use of images from literature and certain websites in this research should constitute fair use and should accordingly not constitute copyright infringement. Our pre-trained model does not generate any visible data similar to the images from these resources, nor is any data from the literature and websites reused in other downstream tasks. Importantly, the pre-trained model is solely for academic research and is not intended for commercial purposes. In summary, the dataset for pretraining RetiZero covers nearly all known fundus diseases, integrating very comprehensive ophthalmic knowledge.

Data for internal domain fundus disease identification

To evaluate the performance of RetiZero in the task of fundus disease identification, we curated three retinal photos datasets across multiple hospitals: 1) Healthcare dataset 1 (H1) where retinal photos were acquired using the Topcon TRC-NW8 and Zeiss VISUCAM-200 devices at the Joint Shantou International Eye Center; 2) Healthcare dataset 2 (H2) where retinal photos was acquired by the Topcon TRC-NW8 and Zeiss VISUCAM-200 devices from four hospitals, including Longchuan People's Hospital, Heyuan; Puning People's Hospital, Jieyang; Wuping Hospital, Longyan; and Pengpai Memorial Hospital, Shanwei; 3) Healthcare dataset 3 (H3) was acquired by the Topcon DRI OCT Triton device at the JSIEC. The H1 and H2 datasets exhibit domain differences due to varying clinic sources, whereas the H3 dataset differs from both H1 and H2 due to the use of distinct fundus images acquisition devices. Clinical assessment and labeling procedure are summarized in Supplementary Fig. 11.

The H1 dataset consisted of 11,414 fundus images, covering 15 categories of fundus diseases and normal condition. We further divided the H1 dataset into training (no. of images = 6942), validation ($n = 2284$), and testing ($n = 2288$) for model fine-tuning, selection, and performance verification, respectively (Supplementary Table 4). The H2 dataset consisted of 7812 fundus images, representing 12 types of fundus diseases and normal condition (Supplementary Table 5). To validate the performance of fine-tuning RetiZero for fundus disease identification on the H2 dataset, we partitioned the H2 dataset into training ($n = 4682$), validation ($n = 1561$), and testing ($n = 1569$) set, respectively, for model fine-tuning, selection, and performance evaluation, respectively. The H3 dataset consisted of 10,863 fundus images across 12 disease categories. It was divided into training ($n = 6511$), validation ($n = 2174$), and testing ($n = 2178$) sets for model fine-tuning, selection, and performance evaluation, respectively (Supplementary Table 6). In this task, RetiZero was fine-tuned for internal domain fundus disease identification using PyTorch on an Nvidia Geforce 3090 GPU (24 G). Adam optimizer and cross-entropy loss function were adopted to guide the model fine-tuning. The total iteration epoch and batch size were set to 100 and 64, respectively.

Data for few-shot fine-tuning

To assess RetiZero's performance in the few-shot fine-tuning, we reorganized the H1, H2, and H3 datasets. Specifically, we randomly selected 5 samples from each disease category of H1, H2, and H3 training set for few-shot fine-tuning, while retaining the validation and testing datasets for model selection and performance evaluation. Detailed information on category and data distribution is provided in Supplementary Tables 7–9. In this task, RetiZero was fine-tuned using PyTorch on an Nvidia Geforce 3090 GPU (24 G). The Adam optimizer and cross-entropy loss function were employed for model optimization, with a total of 1000 epochs and a batch size of 32.

Data for cross-domain fundus disease identification

To evaluate the generality and robustness of RetiZero in cross-domain fundus disease identification, we again reorganized the H1, H2, and H3 datasets. This process identified 11 overlapping diseases categories across the three datasets, which were subsequently renamed as rH1 (no. of images = 10,304), rH2 ($n = 6829$), and rH3 ($n = 10,485$). Details are provided in Supplementary Tables 10–12. We then performed three experimental settings. Specifically, rH1, rH2, and rH3 were sequentially used as internal datasets for model fine-tuning, selection, and internal testing, while the remaining two datasets served as external testing sets to assess. In this task, RetiZero was fine-tuned using PyTorch on an Nvidia Geforce 3090 GPU (24 G). The Adam optimizer and cross-entropy loss function were employed, with the model trained over 100 epochs and a batch size of 64.

Data for the tasks of zero-shot and image-to-image retrieval

We combined the three datasets, H1, H2, and H3 from different hospitals into a dataset named EYE-15, containing 30,089 fundus images that included 14 common fundus diseases and 1 normal category. This dataset was used to validate RetiZero's performance in screening common fundus diseases in zero-shot and image-to-image retrieval tasks. The no. of images in each disease category in EYE-15 was provided in Supplementary Table 1. Moreover, we further collect additional 7007 fundus images acquired by different fundus cameras (such as Topcon TRC-NW8, Zeiss VISUCAM-200, and Topcon DRI OCT Triton) from JSIEC and the other four hospitals (EYE-52 dataset), comprising of 51 fundus diseases and 1 normal condition, to validate the performance of zero-shot fundus disease recognition and fundus disease identification by image-to-image retrieval in a more challenging setting. As shown in Supplementary Table 2, EYE-52 included a wide range of clinically rare fundus diseases, such as albinism, Bietti crystalline dystrophy, choroidal coloboma, and choroidal neoplasm. We adopted Top-1, Top-3, and Top-5 accuracy to evaluate the performance of RetiZero in both tasks of zero-shot fundus disease recognition and fundus disease identification by image-to-image retrieval. Top-K accuracy is a metric used to evaluate the performance of a classification model by determining whether the correct label for a given input appears within the TopK $\{\cdot\}$ predictions made by the model, as follows:

$$\text{Top - K accuracy} = \frac{\sum_{i=1}^N 1\{\hat{y}_i \in \text{TopK}\{\hat{y}_i\}\}}{\text{Total number of inputs } N}, \quad (1)$$

where N and y represent the total number and label of input samples, respectively. TopK $\{\cdot\}$ is the set of the top K predicted labels for the i -th sample, $1\{\cdot\}$ is the indicator function that evaluates to 1 if the ground truth label is in the topK predictions, and 0 otherwise. Supplementary Fig. 11 provides the process of the collection for the EYE-15 and EYE-52 datasets. In this study, we also adopted Precision@1, Precision@3, and Precision@5 as the metrics to evaluate the performance of different foundation models in the task of fundus disease retrieval. Precision@ K is a metric used to evaluate the performance of information retrieval systems and ranking algorithms. It is specifically used to measure the precision of the Top- K results returned by a system. Here is the formula and its explanation:

$$\text{Precision}@K = \frac{|R_{all} \cap R_{Re}@K|}{K} \quad (2)$$

where R_{all} is the set of all relevant samples for the given query, $R_{Re}@K$ represents the set of the Top- K samples retrieved by the system in response to the query. $|R_{all} \cap R_{Re}@K|$ denotes the number of relevant samples in the Top- K retrieved documents, that is the count of samples that are both relevant and retrieved within the Top- K results, while K is the number of Top- samples considered for the calculation.

Framework of RetiZero

Figure 1b provides an overview of the RetiZero framework. RetiZero integrates the advantages of MAE self-supervised learning and CLIP contrastive learning architectures. Specifically, the model was built upon the MAE-based pre-trained backbone network RETFound¹⁵, whose weights were frozen to preserve the model's representation capability for complex semantic information such as lesion contours and topological structures in retinal images. Meanwhile, we introduced low-rank learnable factors into the pre-trained RETFound and leveraged the CLIP architecture to learn image-text knowledge, aiming to enhance the model's understanding of image-text correlations and improve its feature representation capabilities. Furthermore, we incorporated an uncertainty vision-language feature calibration method based on Dirichlet reparameterization into the contrastive

vision-language pretraining framework. This further refined visual-language features in the high-dimensional embedding space, thereby enhancing the model's ability to represent complex features in fundus images. As a result, RetiZero was developed, integrating the advantages of both MAE and CLIP architectures to provide robust feature support for downstream tasks. The components of RetiZero are discussed in details in the following sections.

Image encoder

As shown in Fig. 1b, the image encoder consisted of MAE-based SSL pre-trained backbone and low-rank learnable factors. MAE is a widely used self-supervised learning approach that employs a simple auto-encoder approach to reconstruct the original signal based on partial observations. MAE-based SSL pretraining can guide the network to focus on the rich structural information and contextual features in the images. Therefore, RETFound¹⁵ was adopted as our MAE-based pre-trained backbone. Low-rank learnable factors (LoRA) are a parameter-efficient transfer learning method based on reparameterization³⁵, which utilizes low-rank representations to minimize the number of trainable parameters. It enables a pre-trained large foundation model to incorporate new knowledge into new target tasks, demonstrating robust and state-of-the-art (SOTA) performance in various parameter-efficient transfer learning tasks. Therefore, we utilized low-rank learnable factors to introduce retinal feature description information into the image encoder of RetiZero, enhancing its capacity to represent feature attributes of retinal images. Specifically, given the input token sequence $F_{in} \in R^{Batch \times N \times C_{in}}$ and the output token sequence $F_{out} \in R^{Batch \times N \times C_{out}}$ obtained by the projection layer $W \in R^{C_{out} \times C_{in}}$, LoRA assumes that updates to W should be gradual and stable, where $Batch$ and N represent the batch size and token size of the token sequence, C_{in} and C_{out} are the channel dimension of input token sequence and output token sequence, respectively. Therefore, we applied low-rank approximations to delineate this gradual update. The details of LoRA operation were given in Supplementary Fig. 17. First, freeze the transformer layer to keep W fixed while adding a bypass to complete the low-rank approximation. And, the bypass consists of two linear mapping layers, $A \in R^{r \times C_{in}}$ and $B \in R^{C_{out} \times r}$, where $r \ll \{C_{in}, C_{out}\}$ and was set to 8 in this paper. Thus, the processing of the update layer \hat{W} can be described as:

$$F_{out} = \hat{W}F_{in} \quad (3)$$

$$\hat{W} = W + \nabla W = W + BA \quad (4)$$

The multi-head self-attention mechanism identifies the regions of focus using feature-relevant intensities, and previous works showed that focusing adaptations on Query (Q) and Value (Va) matrices in the transformer's attention mechanism can effectively capture task-specific nuances without the need for extensive retraining or increasing model complexity^{36,37}. Therefore, we apply low-rank approximation to the projection layers of the query and value to influence the attention scores.:

$$\text{Att}(Q, K, Va) = \text{Softmax} \left(\frac{QK^T}{\sqrt{C_{out}}} + B \right) Va \quad (5)$$

$$Q = \hat{W}_Q F = W_Q F + B_Q A_Q F \quad (6)$$

$$K = W_K F \quad (7)$$

$$V = \hat{W}_{Va} F = W_{Va} F + B_{Va} A_{Va} F \quad (8)$$

where K is the Key metric in transformer's attention mechanism, W_Q , W_K , and W_V are frozen projection layers of RETFound, while A_{Va} , B_Q , A_{Va} , and B_{Va} are trainable LORA factors.

Text encoder

Descriptions of fundus images are typically more challenging than those of natural images, as they often contain numerous specialized clinical terms, sometimes even comprising of multiple lesion signs or sentences. Therefore, in this paper, we utilize the BioClinicalBERT³⁸ model as the text encoder for extracting textual features. We use its pre-trained weights on medical texts from the MIMIC III dataset for weight initialization and train it based on a vision-language contrastive learning strategy to extract ophthalmic textual features.

Uncertainty-based feature calibration

Our dataset is collected from various global sources, including public databases, ophthalmic literature, and online resources. Inevitably, it includes complex data distributions such as differences in image resolution, incomplete textual descriptions, and low-quality image-text pairs. Therefore, focusing on these issues, we further introduced an uncertainty vision-language feature calibration method based on Dirichlet reparameterization^{39,40} into the contrastive vision-language pretraining framework, to further calibrate visual-language features in the high-dimensional embedding space for enhancing the robustness of the model to represent complex features in fundus images. Specifically, as shown in Fig. 1b, RetiZero's pretraining consisted of a fundus image encoder and a text encoder. The linear layer serves as a projection head for both the image encoder and the text encoder, mapping the acquired features to a 512-dimensional embedding feature space. Let assume $\phi = \{\phi_E, \phi_H\}$ denotes image encoder (ϕ_E) and corresponding projection head (ϕ_H). Given a fundus image X_i , the image encoder was adopted to obtain feature representation of $F_{Img} \in R^{D_F} : F_{Img} = \phi_E(X_i)$, where D_F is the dimension of image feature space. Meanwhile, $\psi = \{\psi_E, \psi_H\}$ is used to represent text encoder (ψ_E) and corresponding projection head (ψ_H). The text encoder (ψ_E) is adopted to extract feature embedding $F_{Txt} \in R^{D_{Txt}} : F_{Txt} = \psi_E(X_{Txt})$ from text input X_{Txt} , where D_{Txt} is the dimension of text feature embedding. Then, image projection head (ϕ_H) and text projection head (ψ_H) were utilized to map the independent modality representations into a joint unit hyper-sphere space: $V = \frac{\phi_H(F_{Img})}{\|\phi_H(F_{Img})\|}$ and $L = \frac{\psi_H(F_{Txt})}{\|\psi_H(F_{Txt})\|}$, respectively. The similarity between the input image (X_i) and input text (X_{Txt}) are evaluated by the cosine similarity based on the normalized features: $V^T L$, where T represents the transpose operator. With obtained similarity metrics, the optimization goal of the contrastive-based learning pre-training approach is to minimize the distance of features between paired images and text descriptions while maximizing the distance between features of unpaired samples. Specifically, assuming that a batch contains N samples, $V_i \in \{V_1, V_2, \dots, V_N\}$, and $L_i \in \{L_1, L_2, \dots, L_N\}$ represent image feature vector and text feature vector of each sample, while $G = \{0, 1, \dots, N-1\}$ was the corresponding category label, respectively. To guide model optimization, we use the following loss function,

$$\mathcal{L}_{Con} = \mathcal{L}_{Em} + \mathcal{L}_{Dl} \quad (9)$$

where

$$\mathcal{L}_{Em} = -\frac{1}{2N} \sum_{i=1}^N \left(\log \left(\frac{e^{\text{Sim}(V_i^T, L_i)/\tau}}{\sum_{k=1}^N e^{\text{Sim}(V_i^T, L_k)/\tau}} \right) + \log \left(\frac{e^{\text{Sim}(L_i^T, V_i)/\tau}}{\sum_{k=1}^N e^{\text{Sim}(L_i^T, V_k)/\tau}} \right) \right) \quad (10)$$

where N is the samples in a batch, $\text{Sim}(\cdot)$ denotes the cosine similarity function, τ is a trainable scale parameter and is set with reference to FLAIR³⁷. And the details of Dirichlet reparameterization steps was shown in Supplementary Fig. 18, which includes three steps:

Step (1): Obtaining the evidence feature E_{V2L} and E_{L2V} by applying Softplus activation function to similarity metrics between image and text feature embedding to ensure the feature values are larger than 0:

$$E_{V2L} = \text{Softplus}(V^T L) \text{ and } E_{L2V} = \text{Softplus}(L^T V) \quad (11)$$

where $V2L$ and $L2V$ indicate image-to-text and text-to-image contrastive direction.

Step (2): Parameterizing E_{V2L} and E_{L2V} to Dirichlet distribution, as:

$$\begin{aligned} \alpha_{V2L,k} &= E_{V2L,k} + 1, \text{i.e., } \alpha_{V2L,k} = e_{V2L,k} + 1, e_{V2L,k} \\ &= \{\text{Softplus}(V_k^T L_1), \dots, \text{Softplus}(V_k^T L_N)\} \end{aligned} \quad (12)$$

$$\begin{aligned} \alpha_{L2V,k} &= E_{L2V,k} + 1, \text{i.e., } \alpha_{L2V,k} = e_{L2V,k} + 1, e_{L2V,k} \\ &= \{\text{Softplus}(L_k^T V_1), \dots, \text{Softplus}(L_k^T V_N)\} \end{aligned} \quad (13)$$

where $\alpha_{V2L,k}$, $\alpha_{L2V,k}$, $e_{V2L,k}$ and $e_{L2V,k}$ are the k -th contrastive similarity Dirichlet distribution parameters and evidence for the image-text contrastive similarity of the k -th sample in a batch of N samples.

Step (3): Calculating the belief masses and corresponding uncertainty score as:

$$b_{V2L,k} = \frac{e_{V2L,k}}{S_{V2L}} = \frac{\alpha_{V2L,k} - 1}{S_{V2L}}, u_{V2L} = \frac{N}{S_{V2L}} \quad (14)$$

$$b_{L2V,k} = \frac{e_{L2V,k}}{S_{L2V}} = \frac{\alpha_{L2V,k} - 1}{S_{L2V}}, u_{L2V} = \frac{N}{S_{L2V}} \quad (15)$$

where $S_{V2L} = \sum_{k=1}^N (e_{V2L,k} + 1) = \sum_{k=1}^N \alpha_{V2L,k}$ and $S_{L2V} = \sum_{k=1}^N (e_{L2V,k} + 1) = \sum_{k=1}^N \alpha_{L2V,k}$ are the Dirichlet intensities of image-to-text and text-to-image, respectively, used to constrain $1 = \sum_{k=1}^N b_{V2L,k} + u_{V2L}$ and $1 = \sum_{k=1}^N b_{L2V,k} + u_{L2V}$. It can be seen from Eqs. (14) and (15) the probability assigned to k -th sample is proportional to the observed similarity evidence for sample k . Conversely, if less total similarity evidence was obtained, the greater the total uncertainty.

In this study, we associated the Dirichlet distribution with the distribution of feature similarity between images and text descriptions, thereby obtaining belief masses and corresponding overall uncertainty score for the similarity of images and text description for each sample of a batch, based on the evidence collected from the feature similarity matrix. Therefore, we could work out the Dirichlet distribution parameter of $\alpha_{V2L} = [\alpha_{V2L,1}, \dots, \alpha_{V2L,N}]$ and $\alpha_{L2V} = [\alpha_{L2V,1}, \dots, \alpha_{L2V,N}]$ for image-to-text, and text-to-image, while obtaining the multinomial opinions $D(p_{V2L,i}|\alpha_{V2L,i})$ and $D(p_{L2V,i}|\alpha_{L2V,i})$ where $p_{V2L,i}$ and $p_{L2V,i}$ were the sample assignment probabilities on a simplex. Therefore, the loss function for the re-parameterized similarity matrix as follows:

$$\mathcal{L}_{Dl} = \mathcal{L}_{Dl}^{V2L} + \mathcal{L}_{Dl}^{L2V} \quad (16)$$

where,

$$L_{Dl}^{V2L} = L_{Dl-CE}^{V2L} + \lambda^* \mathcal{L}_{KL} \quad (17)$$

$$\mathcal{L}_{DL}^{L2V} = \mathcal{L}_{DL-CE}^{L2V} + \lambda^* \mathcal{L}_{KL} \quad (18)$$

where L_{DL-CE} ($\mathcal{L}_{DL-CE}^{V2L}$ and $\mathcal{L}_{DL-CE}^{L2V}$) was used to ensure that the correct prediction for the sample with highest similarity between image and text yielded more evidence than other samples, while \mathcal{L}_{KL} was used to ensure that incorrect predictions would yield less evidence, and λ was the balance factor that was gradually increased so as to prevent the model from paying too much attention to the KL divergence in the initial stage of training, which might result in a lack of good exploration of the parameter space and cause the network to output a flat uniform distribution.

$$\begin{aligned} \mathcal{L}_{DL-CE} &= \int \left[\sum_{k=1}^N -y_k \log(p_k) \right] \frac{1}{\beta(\alpha_i)} \prod_{k=1}^N p_k^{\alpha_k-1} dp_k \\ &= \sum_{k=1}^N y_k (\psi(S_k) - \psi(\alpha_k)) \end{aligned} \quad (19)$$

where $\psi()$ was the digamma function, while $\beta()$ is the multinomial beta function for the concentration parameter α .

$$\begin{aligned} \mathcal{L}_{KL} &= \log \left(\frac{\Gamma\left(\sum_{k=1}^N \hat{\alpha}_k\right)}{\Gamma(N) \prod_{k=1}^N \Gamma\left(\sum_{k=1}^N \hat{\alpha}_k\right)} \right) \\ &+ \sum_{k=1}^N (\hat{\alpha}_k - 1) \left[\psi(\hat{\alpha}_k) - \psi\left(\sum_{k=1}^N \hat{\alpha}_k\right) \right] \end{aligned} \quad (20)$$

where $\hat{\alpha} = y + (1 - y) \odot \alpha$ is the adjusted parameter of the Dirichlet distribution which could avoid penalizing the evidence of the ground-truth class to 0, and $\Gamma()$ is the gamma function. In general, as shown in Eq. (10) and Eqs. (16)–(20), by introducing higher penalties for uncertain image-txt pairs via uncertainty calibration. For instance, for the text-image feature alignment with high uncertainty, it imposes an extra penalty to avoid the model focusing excessively on incorrect matches during contrastive pre-training.

We performed a comprehensive set of ablation studies (See Supplementary Tables 21–24) to elucidate the individual contributions of each module within RetiZero. The results confirm that our RetiZero model consistently surpasses all other configurations, illustrating the synergistic advantages gained from integrating an MAE-based encoder, text-image contrastive learning, and uncertainty calibration. Furthermore, we conducted additional ablation experiments using only publicly available datasets for pretraining and compared these outcomes to those of our full RetiZero model, which also incorporates data from ophthalmic literature and online resources (See Supplementary Table 25). Although public datasets alone are sufficient for reliably identifying common fundus diseases, they prove inadequate for robustly detecting rare conditions; the inclusion of more diverse data sources is thus indispensable. This finding underscores the critical importance of aggregating varied datasets to develop a knowledge-rich vision-language model with strong clinical applicability across a wide range of diseases.

Definition of Dirichlet distribution

The Dirichlet distribution was parameterized by its concentration K parameters $\alpha = [\alpha_1, \dots, \alpha_K]$ ^{39,40}. Therefore, the probability density function of the Dirichlet distribution was computed as:

$$D(P|\alpha) = \begin{cases} \frac{1}{B(\alpha)} \prod_{k=1}^K p_k^{\alpha_k-1} & \text{for } P \in S_K \\ 0 & \text{Otherwise} \end{cases} \quad (21)$$

where S_K is the K -dimensional unit simplex. and $B(\alpha)$ represents the K -dimensional multinomial beta function.

AI-assisted clinical diagnosis settings

To assess the capability of RetiZero's in recognizing fundus disease recognition without retraining the model, we randomly selected two images from each category of the EYE-52 dataset, creating a subset named EYE52-sub, which containing a total of 104 fundus photographs. We invited 19 ophthalmologists from 12 different institutions and hospitals across Singapore, the United States, and China, to make clinical diagnoses based on these 104 retinal images. Among the participating ophthalmologists, seven have 3–5 years of clinical experience, seven have 5–10 years, and five have more than 10 years. We developed an online fundus image reading system and uploaded the 104 images to the server (Fig. 3). To mimic the zero-shot setup, we provided 52 disease options as prompts on the webpage. During the image reading process, the clinicians selected a diagnosis from these 52 disease categories based solely on the image content. Additionally, each of the ophthalmologists was also asked to rate their confidence in their diagnostic results from level 1 to level 5.

To evaluate whether our RetiZero can assist clinicians in improving their accuracy in diagnosing fundus disease, we conducted a second round of clinical evaluations with the same 19 ophthalmologists, one week after the initial round. We used the same set of retinal images and questions from the first round, but randomized the sequence of the images presented. For each question, RetiZero provided its top five prediction results as references. This approach allows for a comparative analysis of diagnostic performance with and without the assistance of RetiZero's predictions. Moreover, to mitigate potential bias, we conducted additional experiments using a new set of 104 images for the third round of testing to further demonstrate that our RetiZero can assist clinicians in improving their accuracy in diagnosing fundus disease (See Supplementary Table 26).

We also conducted an in-depth analysis of the correlation between the position of the correct diagnosis within RetiZero's top five reference answers and the modifications made by ophthalmologists to their diagnoses. The top five results provided by RetiZero were scored as follows: correct diagnoses appearing in positions 1, 2, 3, 4, and 5 were assigned 5, 4, 3, 2, and 1 points, respectively, while those not appearing in the top five were assigned 0 points. This scoring system allowed us to calculate the top-ranking score for each case and assess how the ranking of the correct diagnosis with in the top five influenced diagnostic accuracy of the ophthalmologists. Moreover, we assessed the modifications made by ophthalmologists to their answers. Assuming an ophthalmologist's answers in the first and second rounds for the same case were denoted as (x, y) , the scoring method was as follows: (True, True) = 0, (False, False) = 0, (True, False) = -1, (False, True) = 1. By analyzing the modification behavior of 19 ophthalmologists, we calculated a response modification score for each case, which reflected the ophthalmologists' thought process and decision-making during the diagnostic process.

Statistics & reproducibility

All results were analyzed and visualized with Python v.3.9, NumPy v.1.26.4, SciPy v.1.11.4, seaborn v.0.13.2, Matplotlib v.3.9.2, pandas v.2.2.2, Scikit-Learn v.1.5.1 and Pillow v.10.4.0. All statistical tests were two-sided, with statistical significance set at a threshold of 0.05. No statistical method was used to predetermine sample size. The sample size was determined based on the availability of the existing dataset to ensure the model could cover a variety of fundus disease types and perform effective training and validation. No data were excluded from the analyses. Each case contains valuable information that is crucial for a comprehensive evaluation of the model's performance. The experiments were randomized. The dataset was randomly divided into training, validation, and test sets to ensure the representativeness of the data the model encountered at different stages and to reduce

potential biases. The Investigators were blinded to allocation during experiments and outcome assessment to avoid subjective influences on the experimental results and improve the objectivity and reliability of the study.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The publicly available datasets used for pre-training are available at the following links and references: APTOS: <https://www.kaggle.com/c/aptos2019-blindness-detection>. Cataract: <https://www.kaggle.com/datasets/jr2nbg/cataractdataset>. DDR: <https://github.com/nkicsl/DDR-dataset>. Diabetic Retinopathy Level Detection: <https://www.kaggle.com/datasets/arbetbi/diabetic-retinopathy-level-detection>. Diabetic Retinopathy Organized: <https://www.kaggle.com/datasets/dola1507108/diabetic-retinopathy-organized>. DR15: https://www.kaggle.com/datasets/nawa393/dr15_test. Messidor: <https://paperswithcode.com/dataset/messidor-1>. MURED: <https://www.kaggle.com/datasets/abhirampoliseti/multi-label-retinal-disease-mured-dataset>. Retina Dataset: <https://www.kaggle.com/datasets/jr2nbg/cataractdataset>. Kaggle DR: <https://www.kaggle.com/c/diabetic-retinopathy-detection/data>. ODIRSK: <https://www.kaggle.com/datasets/andrewmv/ocular-disease-recognition-odirk5>. AUS dataset: <https://www.kaggle.com/competitions/innovation-challenge-2019/data>. SUSTech dataset⁴¹: <https://www.kaggle.com/datasets/mariaherrerot/the-sustechsysu-dataset>. SiMES⁴², SIDNI⁴³, ACRIMA⁴⁴, BEH⁴⁵, DeepDRiD⁴⁶, DRI-2⁴⁷, E-ophta⁴⁸, AIROGS⁴⁹, DeepEyeNet⁵⁰, FIVES⁵¹, G1020⁵², Glaucoma dataset^{53,54}, IDRiD⁵⁵, JICHI⁵⁶, REFUGE⁵⁷, ORIGA⁵⁸, PARAGUAY⁵⁹, EyePACS AirDoc⁶⁰, JSIEC⁹, RFMid⁶¹. Additional data sets supporting the findings of this study are not publicly available due to the confidentiality policy of the Chinese National Health Council and institutional patient privacy regulations. However, they are available from the corresponding authors upon request. For replication of the findings and/or further academic and AI-related research activities, data may be requested from the corresponding author H.C. (drchen-haoyu@gmail.com), and any requests will be responded to within 10 working days. Source data are provided with this paper.

Code availability

All codes are available at <https://github.com/LooKing9218/RetiZero>.

References

1. Bellemo, V. et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. *Lancet Digital Health* **1**, e35–e44 (2019).
2. Xie, Y. et al. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. *Lancet Digital Health* **2**, e240–e249 (2020).
3. Liu, H. et al. Development and validation of a deep learning system to detect glaucomatous optic neuropathy using fundus photographs. *JAMA Ophthalmol.* **137**, 1353–1360 (2019).
4. Wang, M. et al. Characterization of central visual field loss in end-stage glaucoma by unsupervised artificial intelligence. *JAMA Ophthalmol.* **138**, 190–198 (2020).
5. Peng, Y. et al. Automatic staging for retinopathy of prematurity with deep feature fusion and ordinal classification strategy. *IEEE Trans. Med. Imaging* **40**, 1750–1762 (2021).
6. Taylor, S. et al. Monitoring disease progression with a quantitative severity scale for retinopathy of prematurity using deep learning. *JAMA Ophthalmol.* **137**, 1022–1028 (2019).
7. Fauw, D. J. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
8. Wang, M. et al. Uncertainty-inspired open set learning for retinal anomaly identification. *Nat. Commun.* **14**, 6757 (2023).
9. Cen, L. P. et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nat. Commun.* **12**, 4828 (2021).
10. Huang, Z. et al. A visual-language foundation model for pathology image analysis using medical twitter. *Nat. Med.* **29**, 2307–2316 (2023).
11. Radford, A. et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763 (PMLR, 2021).
12. Zhang, M. et al. Asymmetric two-stream architecture for accurate RGB-D saliency detection. In *Computer Vision–ECCV 2020: 16th European Conference, Proceedings, Part XXVIII 16*. Springer International Publishing, 374–390 (ECCV, 2020).
13. Zhang, Z. et al. Few-shot learning for fine-grained signal modulation recognition based on foreground segmentation. *IEEE Trans. Vehicular Technol.* **71**, 2281–2292 (2022).
14. Lai, H. et al. Carzero: Cross-attention alignment for radiology zero-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11137–11146 (IEEE, 2024).
15. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
16. He, K. et al. Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009 (IEEE, 2022).
17. Silva-Rodríguez, J., Chakor, H., Kobbi, R., Dolz, J. & Ben Ayed, I. A Foundation Language-Image Model of the Retina (FLAIR): encoding expert knowledge in text supervision. *Med. Image. Anal.* **99**, 103357 (2025).
18. Selvaraju, R. R. et al. Gradcam: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision*, 618–626 (IEEE, 2017).
19. Quellec, G., Lamard, M., Conze, P. H., Massin, P. & Cochener, B. Automatic detection of rare pathologies in fundus photographs using few-shot learning. *Med. Image Anal.* **61**, 101660 (2020).
20. Gao, M. et al. Discriminative ensemble meta-learning with co-regularization for rare fundus diseases diagnosis. *Med. Image Anal.* **89**, 102884 (2023).
21. Yi, D., Hua, Y., Murchie, P. & Sharma, P. K. Label-free medical image quality evaluation by semantics-aware contrastive learning in IoMT. *IEEE J. Biomed. Health Inf.* **29**, 2335–2344 (2023).
22. Mahapatra, D., Ge, Z. & Reyes, M. Self-supervised generalized zero shot learning for medical image classification using novel interpretable saliency maps. *IEEE Trans. Med. Imaging* **41**, 2443–2456 (2022).
23. Fang, J. et al. Deep metric learning with mirror attention and fine triplet loss for fundus image retrieval in ophthalmology. *Biomed. Signal Process. Control* **80**, 104277 (2023).
24. Fang, J., Fu, H. & Liu, J. Deep triplet hashing network for case-based medical image retrieval. *Med. Image Anal.* **69**, 101981 (2021).
25. Chandakkar, P. S., Venkatesan, R., Li, B. & Li, H. K. A machine-learning approach to retrieving diabetic retinopathy images. In: *Proceedings of the ACM conference on bioinformatics, computational biology and biomedicine*, 588–589 (ACM, 2012).
26. Quellec, G. et al. Automated assessment of diabetic retinopathy severity using content-based image retrieval in multimodal fundus photographs. *Investig. Ophthalmol. Vis. Sci.* **52**, 8342–8348 (2011).
27. Dong, L. et al. Artificial intelligence for screening of multiple retinal and optic nerve diseases. *JAMA Netw. Open* **5**, e229960 (2022).
28. Ruamviboonsuk, P. et al. Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: a

prospective interventional cohort study. *Lancet Digital Health* **4**, e235–e244 (2022).

- 29. Li, B. et al. The performance of a deep learning system in assisting junior ophthalmologists in diagnosing 13 major fundus diseases: a prospective multi-center clinical trial. *npj Digital Med.* **7**, 8 (2024).
- 30. Lin, D. et al. Application of Comprehensive Artificial intelligence Retinal Expert (CARE) system: a national real-world evidence study. *Lancet Digital Health* **3**, e486–e495 (2021).
- 31. Li, B. et al. Development and evaluation of a deep learning model for the detection of multiple fundus diseases based on colour fundus photography. *Br. J. Ophthalmol.* **106**, 1079–1086 (2022).
- 32. Gui, J., Sun, Z., Wen, Y. & Tao, D. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Trans. Knowl. Data Eng.* **35**, 3313–3332 (2021).
- 33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 10684–10695 (IEEE, 2022).
- 34. Shorten, C. & Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *J. Big Data* **6**, 1–48 (2019).
- 35. Hu, E. J. et al. Lora: Low-rank adaptation of large language models. Preprint at <https://arxiv.org/abs/2106.09685> (2021).
- 36. Jie, S. & Deng, Z. H. Fact: Factor-tuning for lightweight adaptation on vision transformer. *Proc. AAAI Conf. Artif. Intell.* **37**, 1060–1068 (2023).
- 37. Chen, C. et al. Ma-sam: Modality-agnostic sam adaptation for 3d medical image segmentation. *Med. Image Anal.* **98**, 103310 (2024).
- 38. Alsentzer, E. et al. Publicly available clinical BERT embeddings. Preprint at <https://arxiv.org/abs/1904.03323> (2019).
- 39. Han, Z., Zhang, C., Fu, H. & Zhou, T. Y. Trusted multi-view classification. In: *International conference on learning representations (ICLR 2020)*.
- 40. Ng, K. W., Tian, G.-L. & Tang, M.-L. *Dirichlet and related distributions: theory, methods and applications* (Wiley, 2011).
- 41. Lin, L. et al. The SUSTech-SYSU dataset for automated exudate detection and diabetic retinopathy grading. *Sci. Data* **7**, 409 (2020).
- 42. Foong, A. W. P. et al. Rationale and methodology for a population-based study of eye diseases in Malay people: the Singapore Malay eye study (SiMES). *Ophthalmic Epidemiol.* **14**, 25–35 (2007).
- 43. Sabanayagam, C. et al. Singapore Indian Eye Study-2: methodology and impact of migration on systemic and eye outcomes. *Clin. Exp. Ophthalmol.* **45**, 779–789 (2017).
- 44. Diaz-Pinto, A. et al. CNNs for automatic glaucoma assessment using fundus images: an extensive validation. *Biomed. Eng.* **18**, 1–19 (2019).
- 45. Islam, M. T. et al. Deep learning-based glaucoma detection with cropped optic cup and disc and blood vessel segmentation. *IEEE Access* **10**, 2828–2841 (2021).
- 46. Liu, R. et al. Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns*, **3**, 100512 (2022).
- 47. Pires, R., Jelinek, H. F., Wainer, J., Valle, E. & Rocha, A. Advancing bag-of-visual-words representations for lesion classification in retinal images. *PLoS One* **9**, e96814 (2014).
- 48. Decenciere, E. et al. TeleOphta: Machine learning and image processing methods for teleophthalmology. *IRBM* **34**, 196–203 (2013).
- 49. De Vente, C. et al. AIROGS: artificial intelligence for robust glaucoma screening challenge. *IEEE Trans. Med. Imaging* **43**, 542–557 (2023).
- 50. Huang, J. H. et al. Deepopty: medical report generation for retinal images via deep models and visual explanation. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2442–2452 (IEEE, 2021).
- 51. Jin, K. et al. Fives: a fundus image dataset for artificial Intelligence based vessel segmentation. *Sci. Data*, **9**, 475 (2022).
- 52. Bajwa, M. N. et al. G1020: a benchmark retinal fundus image dataset for computer-aided glaucoma detection. In: *2020 International Joint Conference on Neural Networks*. 1–7 (IJCNN, 2020).
- 53. Singh, A., Dutta, M. K., ParthaSarathi, M., Uher, V. & Burget, R. Image processing based automatic diagnosis of glaucoma using wavelet features of segmented optic disc from fundus image. *Comput. Methods Prog. Biomed.* **124**, 108–120 (2016).
- 54. Issac, A., Sarathi, M. P. & Dutta, M. K. An adaptive threshold based image processing technique for improved glaucoma detection and classification. *Comput. Methods Prog. Biomed.* **122**, 229–244 (2015).
- 55. Porwal, P. et al. Idrid: diabetic retinopathy—segmentation and grading challenge. *Med. Image Anal.* **59**, 101561 (2020).
- 56. Takahashi, H. et al. Applying artificial intelligence to disease staging: Deep learning for improved staging of diabetic retinopathy. *PLoS One* **12**, e0179790 (2017).
- 57. Orlando, J. I. et al. Refuge challenge: a unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Med. Image Anal.* **59**, 101570 (2020).
- 58. Zhang, Z. et al. Origa-light: an online retinal fundus image database for glaucoma analysis and research. In: *2010 Annual international conference of the IEEE engineering in medicine and biology*, 3065–3068 (IEEE, 2010).
- 59. Benitez, V. E. C. et al. Dataset from fundus images for the study of diabetic retinopathy. *Data Brief.* **36**, 107068 (2021).
- 60. Ju, L. et al. Improving medical images classification with label noise using dual-uncertainty estimation. *IEEE Trans. Med. Imaging* **41**, 1533–1546 (2022).
- 61. Pachade, S. et al. Retinal fundus multi-disease image dataset (rfmid): a dataset for multidisease detection research. *Data* **6**, 14 (2021).

Acknowledgements

H.F. discloses support for the research described in this study from the A*STAR Central Research Fund (“Robust and Trustworthy AI system for Multi-modality Healthcare”). C.-Y.C. discloses support for the research described in this study from the Advanced Manufacturing and Engineering (AME) Programmatic Fund (A20H4b0141) and the National Medical Research Council, Singapore (MOH-CSASI22jul-0001). H. C. discloses support for the research described in this study from Department of Education of Guangdong Province (2024ZDZX2024), Shantou Science and Technology Program (190917085269835), and National Key R&D Program of China (2018YFA0701703). C.Q.Z. discloses support for the research described in this study from the National Natural Science Foundation of China (62376193). C.S.L. discloses support for the research described in this study from NIH (R01AG060942, OT2OD032644).

Author contributions

M.W.: conceptualization, methodology, data collection, experimental deployment, software, writing—original draft. T.L.: data collection, clinical assessment and annotation and curation, software, review and editing. A.L.: clinical support, data annotation and curation. K.Y.: Algorithm design, software, review and editing. Y.P., L.W., C.C., K.Z., Y.M., J.W., C.Q.Z., and X.C.: methodology, writing—review and editing. H.L., and M.C.: data preparing and clinical evaluation, X.Y., M.Z., B.H., C.Z., P.Z., W.C., Y.L., and Y.C.: data preparing. Z.J., M.Q., H.X., T.S., Q.Z., J.G., W.W., S.T., B.J.F., C.S.L., S.Y.W., H.Z., X.L.C., D.F., S.L., and L.S.: clinical evaluation. Q.L., Y.H., Y.C.T., D.L., Y.Z., C.Y.C., C.P.P., and P.A.K.: methodology, writing—review and editing. C.-Y.C.: supervision, clinical assessment and annotation and curation, clinical support, writing—review and editing. H.C.: supervision, clinical assessment and annotation and curation, clinical support, writing—review and editing. H.F.: supervision, project administration, methodology, writing—review and editing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-60577-9>.

Correspondence and requests for materials should be addressed to Ching-Yu Cheng, Haoyu Chen or Huazhu Fu.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025

¹Centre for Innovation and Precision Eye Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. ²Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. ³Joint Shantou International Eye Center, Shantou University and the Chinese University of Hong Kong, Shantou, Guangdong, China. ⁴Department of Radiology, University of Pennsylvania, Philadelphia, PA, USA. ⁵School of Biomedical Engineering, Anhui Medical University, Hefei, Anhui, China. ⁶College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, China. ⁷Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China. ⁸National Key Laboratory of Fundamental Science on Synthetic Vision and the College of Computer Science, Sichuan University, Chengdu, Sichuan, China. ⁹Big Vision Medical Technology Ltd., Suzhou, China. ¹⁰Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore. ¹¹Ophthalmology & Visual Sciences Academic Clinical Program (EYE ACP), Duke-NUS Medical School, Singapore, Singapore. ¹²Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Hong Kong, China. ¹³Shenzhen Longgang E.N.T Hospital, Shenzhen, Guangdong, China. ¹⁴Dongguan Songshan Lake Central Hospital, Dongguan, Guangdong, China. ¹⁵Department of Computer Science, University of Exeter, Exeter, UK. ¹⁶Centre for Medical Image Computing, University College London, London, UK. ¹⁷NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust, London, UK. ¹⁸Institute of Ophthalmology, University College London, London, UK. ¹⁹Tsinghua Medicine of Tsinghua University, Beijing, China. ²⁰School of Clinical Medicine, Beijing Tsinghua Changgung Hospital, Beijing, China. ²¹Foshan Aier Zhuoyue Eye Hospital, Foshan, Guangdong, China. ²²College of Intelligence and Computing, Tianjin University, Tianjin, China. ²³School of Electronics and Information Engineering, Soochow University, Jiangsu, China. ²⁴State Key Laboratory of Radiation Medicine and Protection, Soochow University, Suzhou, China. ²⁵Byers Eye Institute, Department of Ophthalmology, Stanford University School of Medicine, Palo Alto, CA, USA. ²⁶Department of Ophthalmology, University of Washington, Seattle, WA, USA. ²⁷Roger H. and Angie Karalis Johnson Retina Center, Seattle, WA, USA. ²⁸Department of Ophthalmology, Massachusetts Eye and Ear, Harvard Medical School, Boston, MA, USA. ²⁹Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), Singapore, Singapore.

³⁰These authors contributed equally: Meng Wang, Tian Lin.  e-mail: chingyu.cheng@nus.edu.sg; drchenhaoyu@gmail.com; hzfu@ieee.org