npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01772-2

A multimodal visual–language foundation model for computational ophthalmology

Check for updates

Danli Shi[1,2,13] ✉, Weiyi Zhang[1,13], Jiancheng Yang[3], Siyu Huang[4], Xiaolan Chen[1], Pusheng Xu[1], Kai Jin[5], Shan Lin[6], Jin Wei[7], Mayinuer Yusufu[8], Shunming Liu[9], Qing Zhang[10], Zongyuan Ge[11], Xun Xu[7] & Mingguang He[1,2,12] ✉

Early detection of eye diseases is vital for preventing vision loss. Existing ophthalmic artificial intelligence models focus on single modalities, overlooking multi-view information and struggling with rare diseases due to long-tail distributions. We propose EyeCLIP, a multimodal visual-language foundation model trained on 2.77 million ophthalmology images from 11 modalities with partial clinical text. Our novel pretraining strategy combines self-supervised reconstruction, multimodal image contrastive learning, and image-text contrastive learning to capture shared representations across modalities. EyeCLIP demonstrates robust performance across 14 benchmark datasets, excelling in disease classification, visual question answering, and cross-modal retrieval. It also exhibits strong few-shot and zero-shot capabilities, enabling accurate predictions in real-world, long-tail scenarios. EyeCLIP offers significant potential for detecting both ocular and systemic diseases, and bridging gaps in real-world clinical applications.

Ophthalmic diseases such as glaucoma, macular degeneration, and diabetic retinopathy pose a significant threat to global vision health, often leading to vision impairment or even blindness[1]. However, access to timely diagnosis and treatment remains a critical challenge due to insufficient medical resources, especially in underserved regions and developing countries[2,3]. This inequitable distribution of resources makes early detection and intervention for eye diseases particularly challenging, further exacerbating the burden of these diseases.

Computational ophthalmology has emerged as a promising solution, drawing from the concept of "computational pathology."[4,5] This data-driven approach leverages artificial intelligence (AI) and multimodal data to automate image analysis, enhance diagnostic accuracy, and reduce specialists' workloads[6–8]. Recently, the field has shifted from performing specific tasks to developing foundation models[9–14]. After pretraining on a large quantity of labeled or unlabeled data, the model can be easily adapted to downstream tasks in a data-saving manner, reducing the cost and time of data preparation and improving the models' generalization capability. RETFound was the first proposed foundation model in ophthalmology using self-supervised reconstruction learning[10], but it was trained on separate image modalities (color fundus photography [CFP] and optical coherence tomography [OCT]). VisionFM[15] integrates multimodal information through a shared embedding; however, its image encoders remain modality-specific, and a universal model capable of encoding all modalities has yet to be explored. Previously, we proposed EyeFound, which learns a shared representation of multimodal ophthalmic imaging[16]. Nevertheless, existing foundation models still lack modality-modality consistency and image-language alignment—features we consider essential for real-world applications.

In clinical practice, multiple examinations are optimal for examining different eye pathologies, such as CFP, OCT, fundus fluorescein angiography (FFA), and fundus autofluorescence (FAF)[17]. Each examination provides unique and complementary information about the structure and

[1]School of Optometry, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China. [2]Research Centre for SHARP Vision (RCSV), The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China. [3]Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland. [4]School of Computing, Clemson University, Clemson, SC, USA. [5]Department of Ophthalmology, The Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China. [6]Wuhan Bright Eye Hospital, Wuhan, China. [7]Department of Ophthalmology, Shanghai General Hospital, Shanghai Jiao Tong University School of Medicine, National Clinical Research Center for Eye Diseases, No. 100 Haining Road, Shanghai, 20080, PR China. [8]Centre for Eye Research Australia, Royal Victorian Eye and Ear Hospital, East Melbourne, Australia. [9]Department of Ophthalmology, Guangdong Academy of Medical Sciences, Guangdong Provincial People's Hospital, Guangzhou, China. [10]Beijing Tongren Eye Center, Beijing Tongren Hospital, Capital Medical University, Beijing, China. [11]AIM for Health Lab, Faculty of Information Technology, Monash University, Melbourne, VIC, Australia. [12]Centre for Eye and Vision Research (CEVR), 17W Hong Kong Science Park, Science Park, Hong Kong SAR, China. [13]These authors contributed equally: Danli Shi, Weiyi Zhang. ✉e-mail: danli.shi@polyu.edu.hk; mingguang.he@polyu.edu.hk

function of the eye. Previous studies have demonstrated the complementary capabilities of different modalities in enhancing AI models for disease classification and segmentation[18–21]. Therefore, effectively utilizing multimodal data is crucial for obtaining multi-view information, and ensuring consistency across modalities can serve as an important cue for self-supervised learning. Additionally, ophthalmic reports and diagnoses from expert interpretations offer rich textual context, which should be helpful for learning long-tailed representations with hierarchical concepts commonly encountered in the medical domain[11,22]. By integrating clinical text, AI models can better simulate the cognitive processes of human experts, enabling them to handle complex, real-world clinical problems in an ever-changing environment.

In this work, we propose EyeCLIP, an ophthalmic visual-language foundational model designed to harness real-world multi-source, multimodal data. EyeCLIP was pre-trained on a dataset comprising 2,777,593 multimodal ophthalmic images and 11,180 reports from 128,554 patients using self-supervised learning and multimodality alignment. Specifically, the training combined self-supervised reconstruction, multimodal image contrastive, and image-text contrastive learning. Subsequently, we validated EyeCLIP on 14 multi-country datasets to assess its performance in zero-shot, few-shot, and supervised settings across different tasks, including multimodal ocular disease diagnosis and systemic disease prediction, visual question answering (VQA), and cross-modal retrieval. EyeCLIP can effectively learn a shared representation of multiple examinations, enabling zero-shot disease diagnosis and improved language understanding by fully utilizing a large amount of unlabeled, multi-examination, and labeled data in the real world. We believe our approach not only represents a significant advancement in ophthalmic foundation models but also offers insights for training foundational models with incomplete multimodal medical data accumulated in clinical practice across other medical domains.

Results

EyeCLIP development using multi-center multimodal datasets

The EyeCLIP system was trained using 2,777,593 multimodal images and 11,180 reports from 128,554 patients across diverse regions and hospitals in China to learn ophthalmic vision-language features comprehensively. The data details can be found in Fig. 1 and the Methods section. Following training, EyeCLIP can be directly applied to applications involving classification and cross-modal retrieval without further training. Also, it can be finetuned in a data-saving manner for downstream applications such as ocular disease diagnosis, systemic disease prediction, and interactive VQA. Figure 1 shows the study design. The characteristics of the 14 downstream datasets can be found in Supplementary Table 1. Figure 2a presents EyeCLIP's overall superior performance across different downstream tasks compared with the general-domain CLIP[23], medical domain BioMedCLIP[24], PubMedCLIP[25], and the ophthalmology domain RETFound[10].

EyeCLIP excels in zero-shot, partial and full-data training ocular disease classification

Zero-shot transfer capability enables a single pretrained foundation model to be applied directly to downstream tasks. EyeCLIP could be a strong baseline for conventional supervised learning, especially when training labels are scarce. We evaluated EyeCLIP's zero-shot classification performance without task-specific training on nine public ophthalmic datasets. Using CFP as the input modality, EyeCLIP significantly outperformed other models in diagnosing ophthalmic diseases (all $P < 0.001$), with AUCs ranging from 0.681 to 0.757 for DR, 0.721 and 0.684 for glaucoma, as well as 0.660 and 0.688 for multi-disease diagnosis. For OCT, EyeCLIP achieved the highest AUROC scores of 0.800 for OCTID[26] and 0.776 for OCTDL[27], higher than the other models (all $P < 0.001$). Quantitative results are presented in Fig. 2b and Supplementary Table 2.

Next, we evaluated the few-shot performance of EyeCLIP on those nine ocular disease datasets, using limited training samples of 1, 2, 4, 6, and 16, respectively. The results indicated that EyeCLIP could generalize with limited data, demonstrating the ability to diagnose various ophthalmic diseases data-efficiently, outperforming other models (all $P < 0.01$). Quantitative results of AUROC and AUPR are provided in Fig. 3 and Supplementary Table 3.

Specifically, rare diseases are known for lacking sufficient data due to low incidence rates, and they are a common challenge for medical AI, and should be most beneficial with data-efficient training. Therefore, we further evaluated its performance for few-shot classification using a subset of the Retina Image Bank selected by ophthalmologists, with the number of images for each class exceeding 16. The subset included 17 rare diseases: acute posterior multifocal placoid pigment epitheliopathy, birdshot retinochoroidopathy, central areolar choroidal dystrophy, choroidal melanoma, choroidal osteoma, cone dystrophy, congenital hypertrophy of the retinal pigment epithelium, familial exudative vitreoretinopathy, macular telangiectasia, optic disc pit, optic nerve hypoplasia, pseudoxanthoma elasticum, retinitis pigmentosa, retinoblastoma, retinopathy of prematurity, serpiginous choroiditis, Stargardt disease. EyeCLIP beat other models in classifying rare diseases in all settings. The results are presented in Fig. 4c and Supplementary Table 4. Diseases with more distinct clinical features, such as choroidal melanoma and retinitis pigmentosa, were more readily identified across imaging modalities.

Lastly, we tested EyeCLIP using the full-data supervised training paradigm on 11 public datasets containing unimodal and multimodal images, with a train, validation and test split ratio of 55:15:30%. Detailed results are provided in Fig. 4a and Supplementary Table 5.

For single-modality tasks, EyeCLIP outperformed competing models except for three datasets when it was on par with the 2nd best model RETFound. In DR classification, EyeCLIP significantly surpassed RETFound in IDRiD dataset [with AUROC 0.835 vs 0.826, $P = 0.013$], which is a small dataset, but on par with RETFound on much larger datasets APTOS2019 and MESSIDOR2 ($P > 0.05$), suggesting EyeCLIP surpasses RETFound in a matter of data efficiency, requiring less data than RETFound. For glaucoma and multi-disease classification, EyeCLIP consistently outperformed other models. For OCT images, EyeCLIP was on par with RETFound on OCTID dataset ($P > 0.05$), but significantly better on OCTDL dataset (AUROC 0.993 vs. 0.982, $P < 0.001$), which is a more imbalanced dataset with long-tailed classes. Even though RETFound specifically trained separate weights that are optimal for CFP and OCT, EyeCLIP is generally better and no worse than it, even with a single general encoder.

For multimodality tasks, EyeCLIP outperformed all comparison models. On the AngioReport (APTOS2023[28]) dataset with two modalities, EyeCLIP outperformed the next best model, BioMedCLIP, with an AUROC of 0.721 versus 0.705, $P < 0.001$. Moreover, EyeCLIP performed the best on the challenging Retina Image Bank[29] dataset with 14 modalities and 84 conditions, including rare diseases, with AUROC of 0.561 versus the 2nd best 0.545, $P < 0.001$.

We also conducted an ablation study on fully supervised training to investigate the contributions of image-text contrastive learning, image-image contrastive learning, and image self-reconstruction learning. Results can be found in Supplementary Table 6 and Supplementary Fig. 2. The pretraining and downstream settings in the ablation experiments remained consistent with the original EyeCLIP model. The results indicate that removing any of these components leads to a decline in performance on downstream tasks, demonstrating the necessity and effectiveness of the EyeCLIP design. Among them, the model without image self-reconstruction learning experienced the most significant performance drop. For instance, on the multimodal dataset Retina Image Bank, the AUROC decreased by 14.2 percentage points, while on the AngioReport dataset, it dropped by 14.6 percentage points. This suggests that image self-reconstruction learning plays a crucial role in maintaining robust feature representations, particularly in scenarios with diverse modalities, where reconstructing input images helps preserve structural and semantic consistency across different imaging techniques.

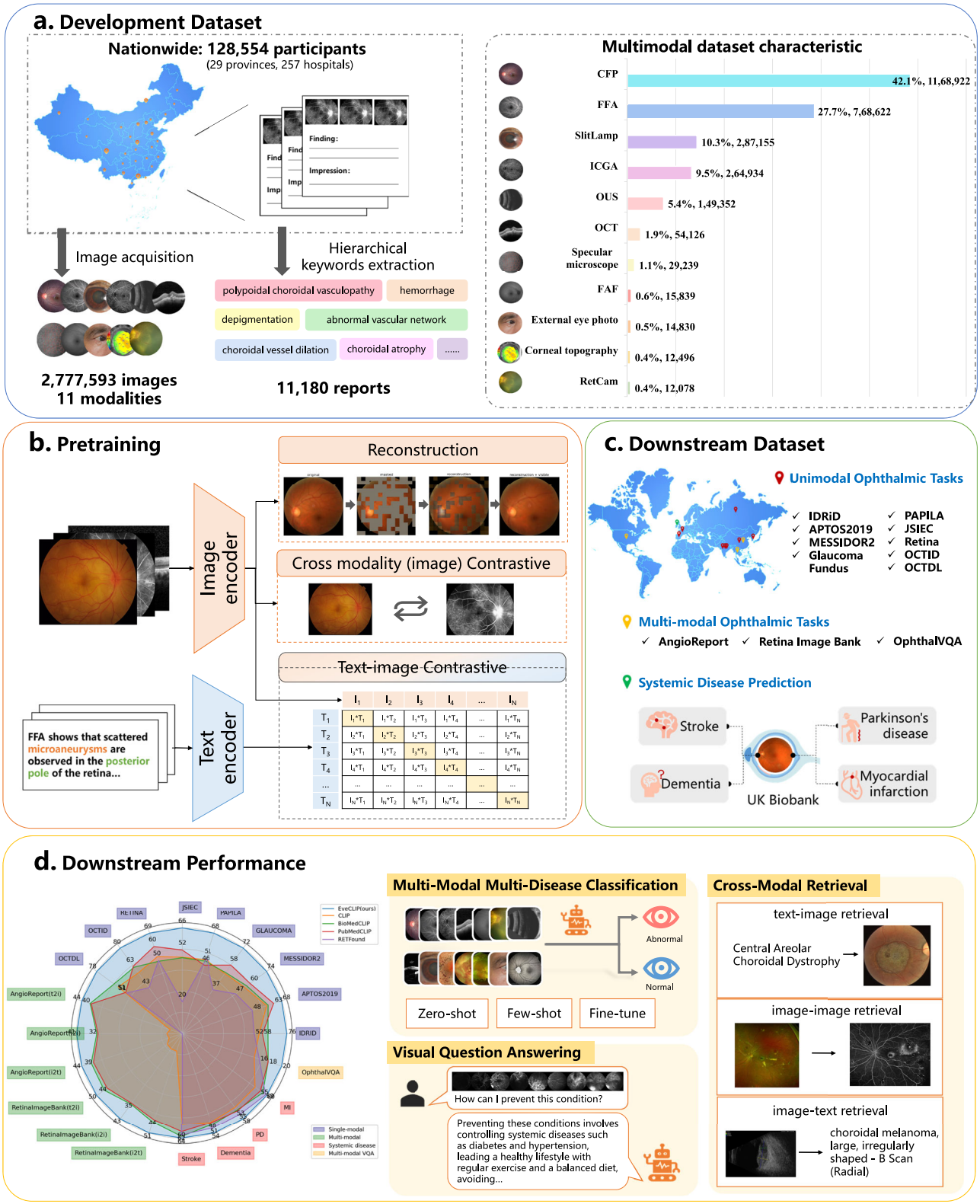

Fig. 1 | Study diagram. a Using an extensive multimodal database across nine provinces in China, we matched the multi-examination images from the same patient, and cleaned the medical reports using a keyword mapping dictionary containing medical terminology to generate hierarchical keyword text labels. **b** EyeCLIP was pretrained using self-supervised reconstruction, multi-examination contrastive learning, and hierarchical text-image contrastive learning to leverage real-world multi-examination clinical data fully. **c** Downstream multi-country datasets for EyeCLIP validation, including zero-shot, few-shot, and supervised finetuning scenarios. **d** Radar plot outlines the performance of EyeCLIP and baseline models across various downstream tasks. EyeCLIP significantly outperforms the baseline models across diverse tasks, including zero-shot classification, multimodal retrieval, visual question answering (VQA), and supervised systemic disease prediction.

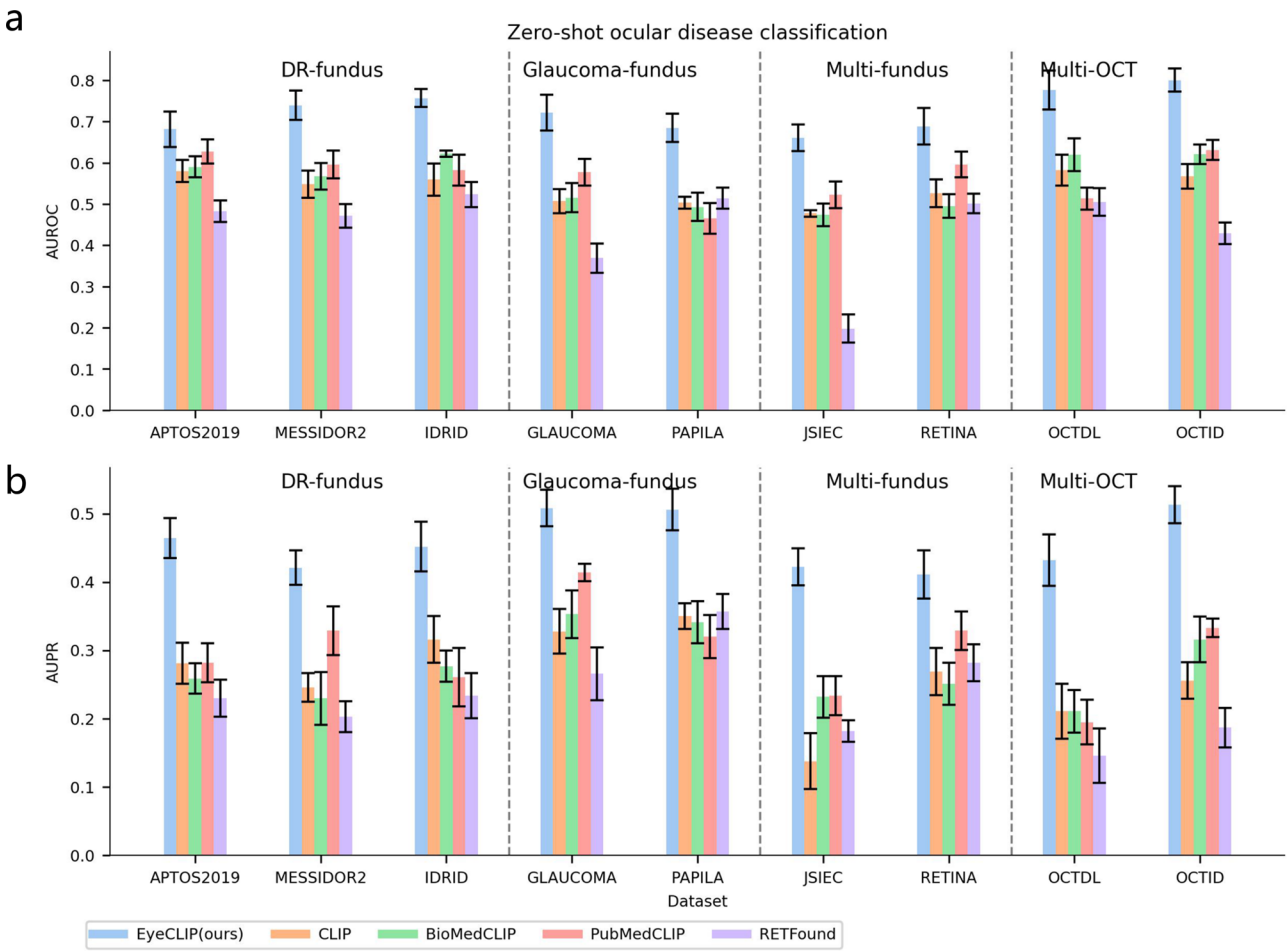

Fig. 2 | Zero-shot performance on downstream ocular diseases datasets. **a** AUROC. **b** AUPR. Error bars represent 95% confidence intervals, and the centers correspond to computed values of each metric. EyeCLIP achieved significantly better zero-shot performance than other models for both AUROC and AUPR. AUROC = area under the receiver operator characteristic curve, AUPR = area under the precision-recall curve. EyeCLIP outperforms the second-best model FLAIR, a pretrained vision-language model for universal retinal fundus image understanding. Notably, FLAIR was pretrained on public datasets, with its performance evaluated through internal validation. In contrast, EyeCLIP, which was not trained on these public datasets, demonstrated its performance through external validation, highlighting its strong generalizability.

EyeCLIP enhances systemic disease prediction

Systemic diseases such as stroke and myocardial infarction (MI) pose significant threats to older adults, often leading to sudden death. The eyes, rich in blood vessels that can be directly visualized, have been referred to as "the window to the body's health."[30,31] Therefore, predicting the incidence of systemic diseases is a crucial technique for early screening and prevention. However, compared to the general population, the incidence of these events is relatively low, resulting in limited positive training data. Consequently, data-efficient training methods are highly valued in this context. We evaluated EyeCLIP's performance in predicting systemic diseases based on ophthalmic images using the UK Biobank[32]. Our experiment included predictions for stroke, dementia, Parkinson's disease (PD), and MI. We first assessed the few-shot performance of EyeCLIP using limited training samples of 1, 2, 4, 8, and 16, respectively. EyeCLIP consistently outperformed other models, demonstrating superior data efficiency in predicting systemic diseases. For full-data supervised training, EyeCLIP ranked first, achieving AUROC scores of 0.641, 0.536, 0.580, and 0.596, and AUPR scores of 0.627, 0.572, 0.616, and 0.582, respectively (all $P < 0.05$). Detailed results are provided in Fig. 4b and Supplementary Tables 7-8.

EyeCLIP achieves zero-shot cross-modal retrieval

By learning an aligned latent space for multimodal embeddings, EyeCLIP enabled zero-shot cross-modal retrieval. This included retrieving text entries based on image queries (image-to-text, i2t), retrieving images based on text queries (text-to-image, t2i), and retrieving images based on image queries (image-to-image, i2i). This function is useful for biomedical applications such as identifying cases for research cohorts, assisting with rare disease presentations, and creating educational resources. We evaluated EyeCLIP on two external multimodal image-caption datasets, AngioReport and Retina Image Bank, which cover a diverse range of ophthalmology concepts. To specifically investigate the performance on rare diseases, we manually selected a subset from Retina Image Bank containing only rare diseases. Following previous studies[11,33], we used Recall@K as the metric for cross-modal retrieval.

On AngioReport, EyeCLIP achieved mean recall of 44.1%, 40.7%, and 44.3% for text-to-image, image-to-image, and image-to-text retrieval, respectively, outperforming BioMedCLIP's 40.5%, 32.9%, and 40.1% ($P < 0.01$ for all tasks). On Retina Image Bank, EyeCLIP achieved a mean recall of 50.2%, 43.3%, and 50.9%, outperforming BioMedCLIP's 45.8%, 35.8%, and 45.3% ($P < 0.01$ for all tasks). Supplementary Table 9 provides details on the model's performance. Examples of the retrieved results are presented in Fig. 5; EyeCLIP effectively retrieved similar contents using text or images as queries. It could retrieve relevant images based on text descriptions, pair images with the same pathological condition or from the same patient, and find the most correlated description with the image inputs.

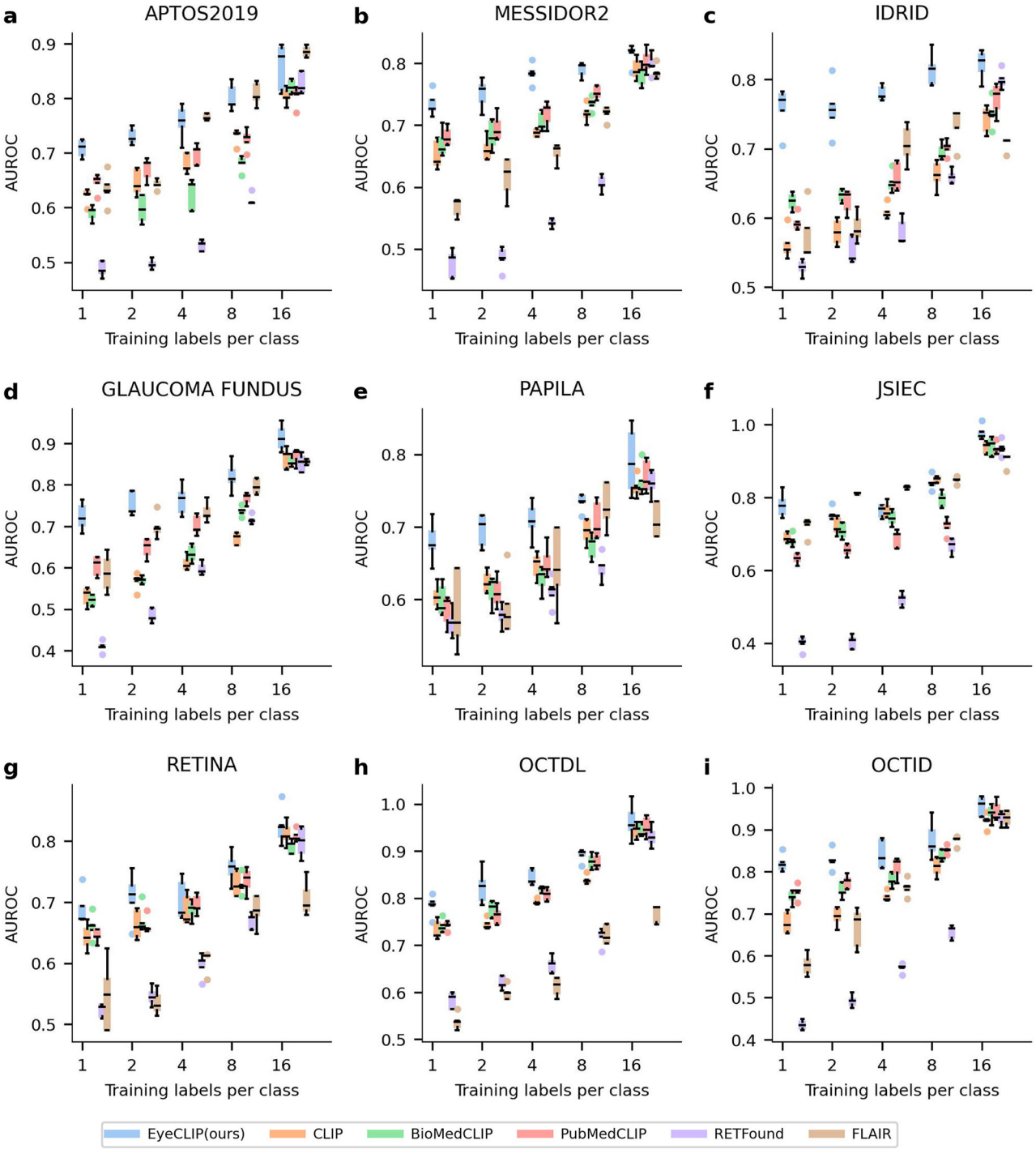

Fig. 3 | Few-shot classification experiments. We investigated the label efficiency of different pretrained models in a few-shot setting, varying the number of training labels per class (nc = 1, 2, 4, 8, 16) in the APTOS2019 (**a**), MESSIDOR2 (**b**), IDRID (**c**), GLAUCOMA FUNDUS (**d**), PAPILA (**e**), JSIEC (**f**), RETINA (**g**), OCTDL (**h**), and OCTID (**i**) dataset. For each nc, we sampled five different sets of training examples and trained a weakly supervised model. Boxes indicate quartile values, and whiskers extend to data points within 1.5× the interquartile range. EyeCLIP achieves significantly better performance (in terms of the mean AUROC of five runs) than other encoders for different sizes of training sets and across all datasets. AUROC = area under the receiver operator characteristic curve. AUPR results can be found in Supplementary Fig. 1.

EyeCLIP demonstrates few-shot generalization on VQA

Vision-language foundation models have great potential for generalization in ophthalmic VQA. We combined the image encoder from each model with a text encoder using a large language model (LLM), specifically Llama2-7b, to perform VQA. We conducted the few-shot VQA on the OphthalVQA[34] dataset. OphthalVQA is an open-set VQA dataset, with the training set comprising 7,778 images across six modalities and VQA pairs covering 40 diseases, including rare conditions. The test set includes 60 images across the same six modalities, representing 60 ophthalmic conditions and 600 QA pairs. As detailed in Supplementary Table 10, EyeCLIP demonstrated superior alignment with the LLM, despite the image and language modules being fine-tuned on a limited amount of VQA data. EyeCLIP ranked first in terms of exact matching score and F1 score across all settings with support numbers of 1, 2, 4, and 8

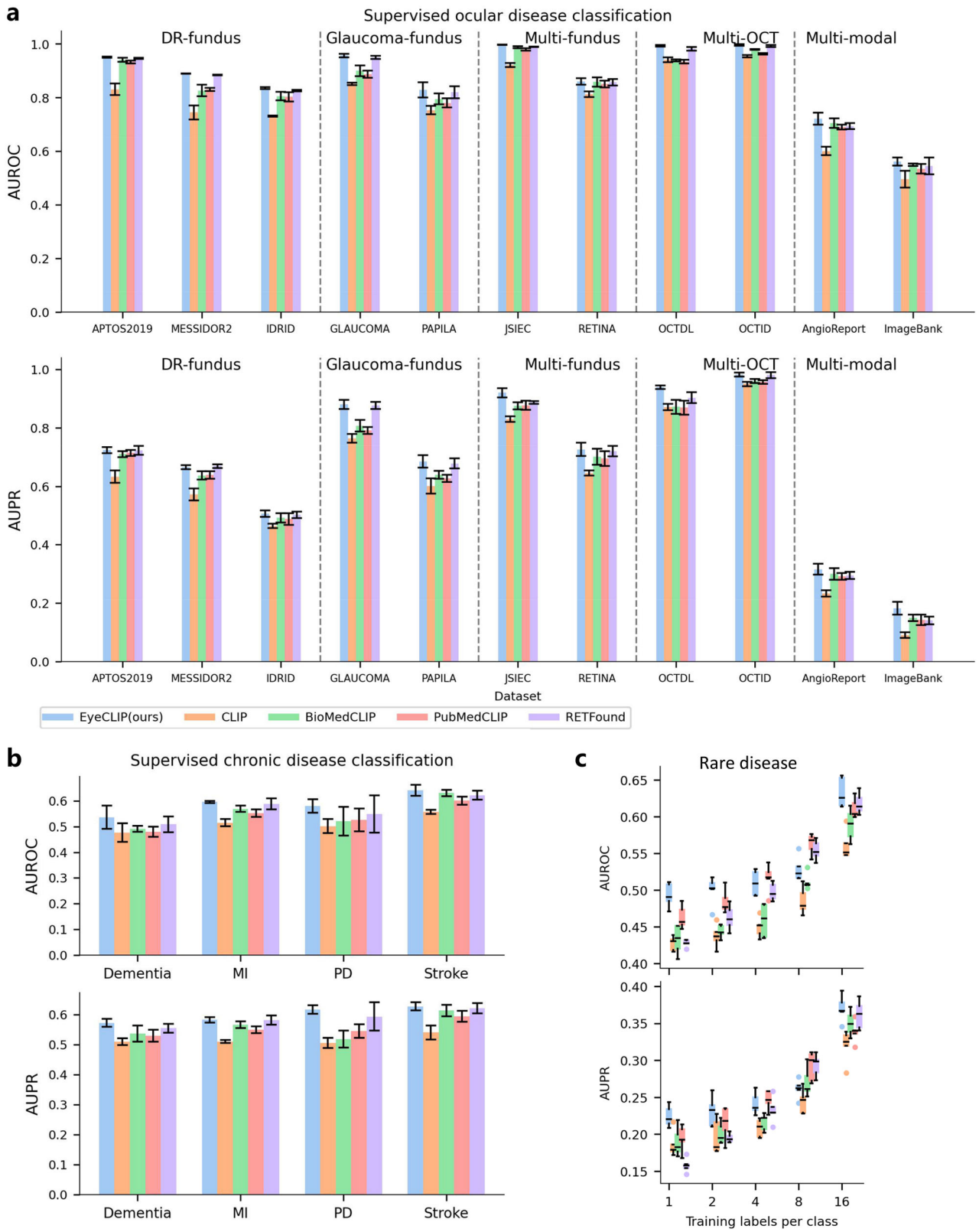

Fig. 4 | Performance of EyeCLIP across ocular, systemic, and rare disease prediction tasks. a Supervised full-data finetuning on ocular disease tasks. EyeCLIP is on par with the 2nd best model RETFound on APTOS2019, MESSIDOR2, OCTID ($P > 0.05$), and surpasses all models on the other eight datasets. **b** Supervised full-data finetuning on systemic disease prediction. EyeCLIP surpasses all other models. ($P < 0.05$). **c** Few-shot finetuning on rare disease classification. EyeCLIP surpasses all other models. ($P < 0.05$). Boxes indicate quartile values, and whiskers extend to data points within 1.5× the interquartile range. Detailed statistics can be found in Supplementary Tables 4-5. AUROC = area under the receiver operator characteristic curve, *AUPR* area under the precision-recall curve.

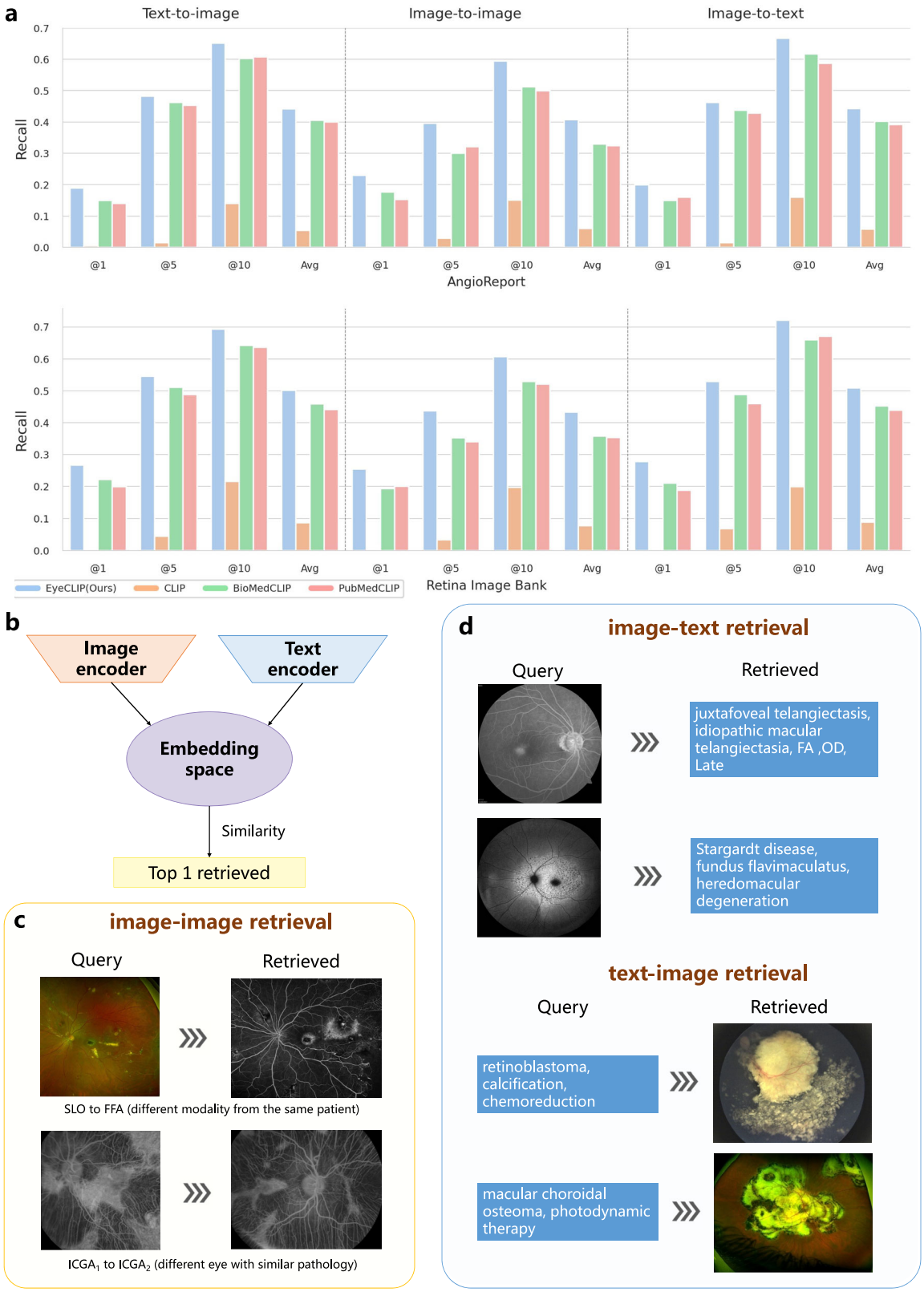

Fig. 5 | Zero-shot multimodal retrieval performance. a Model comparison on two datasets with image-text pairs, AngioReport and Retina Image Bank. Similarity in the embedding space was computed between the query image and all text samples in the database. The top-K most similar texts were retrieved. We report Recall@K for K ∈ {1, 5, 10} and the mean recall, which averages over K. We compared different models in text-to-image (1st column), image-to-image (2nd column) and image-to-text (3rd column). EyeCLIP outperforms other baselines on all retrieval tasks. Error bars indicate 95% confidence intervals. **b** Schematic illustrates zero-shot cross-modal retrieval. **c**, **d** Examples of images in the top one retrieved result from the Retina Image Bank. More examples can be found in Supplementary Fig. 3.

EyeCLIP reveals disease-relevant regions via patch-level similarity visualization

To investigate which image regions the model attends to under different disease conditions, we performed a patch-wise similarity analysis using EyeCLIP. Given a disease-specific textual prompt (e.g., "Color fundus photography, diabetic retinopathy"), we calculated the cosine similarity between the normalized text embedding and each visual patch token obtained from the image encoder. The resulting similarity maps were visualized as heatmaps overlaid on the original fundus images. As illustrated in Supplementary Fig. 4, the model consistently highlights clinically relevant regions, such as the cherry-red spot and retinal edema in central retinal artery occlusion, enlarged optic disc cup in glaucoma, abnormal hyper-/hypo-fluorescence on FFA in wet age-related macular degeneration and choroidal melanoma, hyperreflective bands in retinal detachment, hyperreflective bumps in drusen, corneal ulcers, and abnormal conjunctival features in conjunctivitis. These results suggest that the model is capable of semantically aligning textual disease descriptions with spatially meaningful features in retinal images.

Discussion

In this study, we developed EyeCLIP, a visual-language foundation model for multimodal ophthalmic image analysis, utilizing a large dataset of 2,777,593 ophthalmic images spanning 11 modalities, along with corresponding hierarchical language data. Our novel training strategy fully leverages real-world data nature, characterized by multi-examination and large amounts of unlabeled and labeled data. This approach achieved a shared representation across multiple examinations and modalities. EyeCLIP significantly enhances the analysis of ophthalmic and systemic diseases, demonstrating state-of-the-art efficiency and generalizability in zero-shot, few-shot, and full-data finetuning downstream tasks.

One primary advantage of EyeCLIP lies in its alignment of multiple examinations, which is demonstrated in the image-image retrieval task and multimodal image classification tasks. In contrast, conventional foundation models often focus on specific types of examination, which limits their effectiveness for real-world applications. Given the complexity of real-world clinical settings, where patients present with various conditions and undergo multiple tests, a model capable of accurately identifying diverse eye conditions with different image modalities is highly desirable. Our ablation study further highlights the necessity of EyeCLIP's hybrid approach. Models relying solely on contrastive learning (e.g., CLIP, BioMedCLIP, PubMedCLIP, and FLAIR) without image self-reconstruction learning suffer significant performance drops. This suggests that image self-reconstruction learning is crucial for maintaining robust feature representations, particularly when handling diverse modalities, as it helps preserve structural and semantic consistency across different imaging techniques. Conversely, using only self-reconstruction without contrastive learning (e.g., RetFound) also leads to performance degradation, indicating that contrastive learning is essential for aligning cross-modal features and enhancing the model's ability to leverage complementary information. Compared to VisionFM, which adopts modality-specific encoders with a shared embedding space, EyeCLIP employs a unified encoder to process various ophthalmic modalities. This design offers better scalability and reduces the burden of training and deploying multiple encoders, making it more practical for real-world clinical deployment. While modality-specific designs like RETFound and VisionFM may capture finer modality-specific details, they add architectural complexity and limit adaptability when new or low-resource modalities are introduced. EyeCLIP's single-encoder framework promotes flexible modality alignment and streamlines integration into multi-exam clinical workflows.

EyeCLIP was developed using 11 imaging modalities collected from diverse populations, making it uniquely powerful for diagnosing vision-threatening diseases, particularly in multimodal, multi-disease diagnostics with label imbalance and rare diseases. Notably, the challenging Retina Image Bank underscores its potential for managing rare eye conditions with diverse examination. This capability likely arises from its cross-modal representation learning during pretraining, enabling the capture of complementary patterns across various imaging modalities. However, to establish explicit causal links between these multimodal interactions and model performance, future studies incorporating specialized interpretability frameworks are needed[35].

Another major strength of EyeCLIP is its easy integration into the visual-language framework. While previous foundation models primarily focused on extracting meaningful patterns from rich image data, EyeCLIP utilized textual descriptions created by medical professionals to distill hierarchical context information. By employing text-image contrastive learning, EyeCLIP maximized the use of all available labeled ophthalmic data, learning semantically rich features that align visual patterns with clinical concepts. This alignment offers zero-shot capabilities, significantly reducing the need for extensive annotation of training data. When integrated with LLM, its few-shot VQA capability presents a unique opportunity to automate interpretative tasks in clinical settings with minimal model adjustments and training data. Highlighting the possibility of integrating EyeCLIP into the existing ophthalmic chat systems to perform multitask VQA[8,36–38]. EyeCLIP's ability to operate with minimal training data and adapt to new tasks makes it a valuable tool for expanding the reach of quality ophthalmic care widely.

Ophthalmic images are increasingly used to indicate systemic diseases due to their accessibility[39–42]. This is where the foundation model could be well appreciated due to the scarcity of event data compared with a healthy population. Notably, EyeCLIP significantly improved systemic disease prediction, surpassing previous medical domain foundation models, such as BioMedCLIP and the ophthalmology domain model RETFound, in events including stroke, dementia, PD, and MI. This improvement is likely attributed to the shared representation of different examination data. For example, angiography provides better visualization of retinal blood vessels and lesions, and these features could be jointly learned by the model. After further optimization, EyeCLIP can be a powerful tool for early detection and monitoring of systemic diseases, enhancing patient care beyond ophthalmology. Future studies could further improve predictive performance by integrating ocular biomarkers with multimodal systemic health data, including electronic health records.

This study offers insights for other medical domains dealing with incomplete or unaligned data. In real-world clinical practice, it is common for datasets to contain multimodal information, such as images and text, that are not fully aligned across every sample. In this work, we address this challenge by employing a strategy that combines self-supervised learning through masked-image reconstruction within single modalities and contrastive learning across aligned multimodal data when available. This approach maximizes the utility of diverse clinical data accumulated in practice, offering a potential framework for developing medical foundation models in other fields where incomplete multimodal data is prevalent.

Our study has several limitations. Firstly, EyeCLIP's performance relies on the quality and diversity of the training data. For example, while it shows robust diagnostic capabilities in cross-population CFP classification, a performance discrepancy exists across ethnic groups, with higher AUC in East Asian cohorts (JSIEC 0.977; Glaucoma Fundus 0.913), likely due to the Chinese-dominated training data. To address this bias and enhance generalizability, future work should incorporate more balanced and ethnically diverse datasets, covering underrepresented populations. Moreover, strategies such as demographic-aware sampling, domain adaptation, and cross-domain contrastive learning will be explored to mitigate population-specific biases and improve the model's fairness and reliability across diverse clinical settings. Secondly, many modalities in our dataset, such as FFA, ICGA, and OCT, are inherently 3D, capturing essential dynamic lesion changes and volumetric information[43,44]. However, in this version, we utilized only 2D slices. Future work incorporating the full 3D information may further enhance the model's performance and capability. Thirdly, while EyeCLIP requires no more than 8 GB for single-image inference and is deployable on common edge GPUs (with detailed comparisons of inference time and

training resource requirements provided in Supplementary Table 11), real-time clinical use may benefit from model distillation and quantization to further reduce computational demands[45]. Additionally, ensuring interpretability and transparency is crucial for gaining the trust of healthcare providers and patients, ensuring successful implementation in clinical practice.

In conclusion, we developed EyeCLIP, a visual-language foundation model characterized by shared multimodal representations capable of performing a wide range of downstream tasks. The novel training strategy aligns well with real-world data characteristics, potentially informing the development of foundation models in general medicine. EyeCLIP's outstanding performance and broad applicability to ocular and systemic diseases position it as a promising tool to enhance the accuracy, efficiency, and accessibility of AI in ophthalmic clinical practice and research.

Methods

Ethics statement

This study was conducted in accordance with the Declaration of Helsinki and received approval from the Hong Kong Polytechnic University's institutional review board (HSEARS20240202004). The Institutional Review Board waived informed consent due to the retrospective analysis of anonymized ophthalmic images and public datasets.

Data curation and preprocessing for pretraining

We collected a vast amount of unlabeled ophthalmic images from 227 hospitals across China, totaling 2,777,593 images from 128,554 participants. The gender distribution was balanced (68,531 female, 59,994 males, and 29 unknown). The participants had a mean age of 50.4 years (SD: 23.3, range: 1–98), representing a broad demographic. All participants were of Chinese ethnicity. These images covered a variety of ocular conditions and comprised 11 different image modalities, including CFP, FFA, indocyanine green angiography (ICGA), and OCT, among others. Not all patients underwent imaging for all 11 modalities, leading to missing modalities for some individuals. The percentage of images per modality and class distribution can be found in Fig. 1a, "Multimodal dataset characteristic."

To ensure the quality of the data, we excluded low-quality images from CFP, FFA, and ICGA by extracting and analyzing the vascular structures[40,46]. Specifically, images with detachable vascular ratios less than 0.04 for CFP and less than 0.01 for FFA and ICGA were removed. Images from other modalities were sampled (50 images per modality) and manually reviewed as of sufficient quality. Additionally, since they were captured in clinical settings and optimized for patient distribution, no specific quality control method was applied. The language training data were sourced from 11,180 angiography reports of 11,180 participants. Since the reports contain custom templates and are generally lengthy, we developed a custom dictionary by integrating medical expert knowledge using keyword-based regular expressions to extract essential medical concepts from the report texts. The dictionary includes tree-structured keywords [e.g., "Diabetic retinopathy (DR) → Non-proliferative DR (NPDR) → Mild NPDR"], where parent nodes represent broader concepts and child nodes correspond to more specific terms. The medical reports were therefore converted into a set of keywords covering various aspects such as ophthalmic diseases, anatomical structures, and diagnostic indicators[8,36]. This process provided crucial semantic information for subsequent image-text alignment and pretraining. Before model development, all data, including images and ophthalmic reports, were de-identified. Additional information about the pretraining dataset is summarized in Fig. 1.

To facilitate multimodal alignment, we matched ophthalmic images from different examinations to obtain image pairs from the same patient, enabling the model to learn features across different imaging modalities more effectively.

Data curation and preprocessing for downstream validation

Supplementary Table 1 summarizes the details of datasets used for downstream validation. We included 14 datasets, covering ocular disease diagnosis (single-modality classification, multimodality classification, and VQA) and systemic disease prediction.

Ophthalmic single-modality classification datasets. We compiled 9 publicly available single-modality ophthalmic disease classification datasets from diverse ethnicities and regions, comprising 7 CFP and 2 OCT datasets. The CFP datasets included IDRiD (India, 516 images)[47], APTOS2019 (India, 3662 images), and MESSIDOR2 (France, 1744 images) for DR diagnosis; PAPILA (Spain, 488 images)[48] and Glaucoma Fundus (South Korea, 1544 images)[49] for glaucoma diagnosis; as well as JSIEC[50] and Retina for the classification of multiple ophthalmic diseases. The OCT datasets included OCTID (India, 572 images)[26] and OCTDL (Russia, 2064 images)[27], both containing multiple disease labels.

Ophthalmic multimodality classification datasets. We also collected two multimodality, multi-label datasets: the AngioReport[28] dataset and the Retina Image Bank[29]. The AngioReport dataset comprises approximately 50,000 angiographic images collected from routine clinics in Thailand, encompassing FFA and ICGA modalities and covering 142 retinal diseases. We selected a test subset of 10,520 images to validate our model. The Retina Image Bank, sourced from the United States, is a large open-access repository of retinal images containing 14 modalities and 84 ophthalmic diseases. We obtained images and their corresponding findings from the website and created a custom dictionary to standardize different disease expressions using keyword matching and regular expressions. The standardized labels incorporate hierarchical structures, such as "DR, mild DR" for mild diabetic retinopathy. We excluded non-standard retinal examination images, including schematic cartoons, histology, and pathology images. To increase efficiency, we focused on images uploaded between 2019 and 2023 and removed instances with fewer than 50 occurrences. This process yielded a final dataset of 3293 images.

Ophthalmic VQA Dataset. OphthalVQA(test)[34] is a dataset of ocular multimodal images from China, including CFP, OCT, FFA, slit-lamp, scanning laser ophthalmoscopy (SLO), and ocular ultrasound images. Ten images representing distinct diagnoses were selected for each modality, resulting in a test set of 60 images and 600 free-form question-answer (QA) pairs generated by ophthalmologists. These images reflect typical disease manifestations commonly used for clinical diagnosis. This dataset serves as the testing dataset for the VQA downstream task.

To facilitate few-shot VQA experiments, we manually curated a training dataset, OphthalVQA(train), aligned with the OphthalVQA format. This training dataset comprises five modalities—CFP, FFA, OCT, B-scan ultrasound, and slit lamp—encompassing 54 diseases or conditions. It includes 7778 open-ended QA pairs, with each disease or condition represented by 7 to 16 images. Each image is paired with 6 to 13 questions covering imaging modalities, laterality, diagnosis, image descriptions, and lesion-specific inquiries. Supplementary Table 1 provides detailed characteristics of the dataset.

Systemic chronic disease dataset. UK Biobank[32] is a population-based prospective cohort from the United Kingdom, recruiting approximately 500,000 participants aged 40 to 69 between 2006 and 2010. Among these participants, 82,885 underwent CFP examinations, generating a dataset of 171,500 retinal images. To define outcomes, we used algorithm-based classifications (Category 42) developed by the UK Biobank outcome adjudication group. These algorithms integrate coded information from baseline assessments and linked datasets, providing replicable outcome definitions for major systemic diseases, including stroke, dementia, Parkinson's disease (PD), and myocardial infarction (MI). This method eliminates the need to manually select diagnostic and procedural codes, ensuring reliable and reproducible outcome definitions. To minimize potential biases from variations in individual visits, we included only the retinal images of the right eye from a single visit per patient.

Model Design and Training Details

All experiments were conducted in Python 3.10. For visual-language pretraining, we employed CLIP[23] as our base framework, which is a pretrained

model that leverages contrastive learning using natural image-text pairs. This model processes image and text inputs independently through an image encoder and a text encoder, generating distinctive vector representations for each modality. Subsequently, these vectors are projected into a unified multimodal embedding space, facilitating direct comparisons between textual and visual elements.

We extended the traditional CLIP architecture by adding an image decoder to the CLIP image encoder, following Masked Autoencoders (MAE)[51]. This addition enables the model to perform masked image reconstruction, which is pivotal for self-supervised feature representation learning. Specifically, besides the original image-text contrastive loss $\mathcal{L}_{img-text}$, we modified the loss function of CLIP by adding an image reconstruction loss $\mathcal{L}_{recon}$, and an image-image contrastive loss $\mathcal{L}_{img-img}$.

$\mathcal{L}_{img-text}$ is used to align the image and corresponding text descriptions, which is defined as:

$$\mathcal{L}_{img-text} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\mathrm{sim}\left(f\left(x_i\right),g\left(t_i\right)\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\mathrm{sim}\left(f\left(x_i\right),g\left(t_j\right)\right)/\tau\right)} \quad (1)$$

where *f(x)* and *g(t)* are the encoded image and text representations, *sim* denotes the similarity measure, typically cosine similarity, and τ is a temperature parameter. τ controls the sharpness of the similarity distribution in contrastive learning. A lower τ enhances discrimination but may cause training instability, while a higher τ smooths the distribution but weakens hard negative differentiation. Empirical tuning determined $\tau = 0.07$ as optimal for stable and effective alignment.

Similarly, $\mathcal{L}_{img-img}$ aligns the features between different modalities of images, which is defined as:

$$\mathcal{L}_{img-img} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\mathrm{sim}\left(f\left(x_i\right),f\left(x_j\right)\right)/\tau\right)}{\sum_{k=1}^{N}\exp\left(\mathrm{sim}\left(f\left(x_i\right),f\left(x_k\right)\right)/\tau\right)} \quad (2)$$

To align different image modalities, we employ a shared vision encoder that processes all modalities under a unified contrastive learning objective, encouraging the model to learn modality-invariant embeddings without explicit fusion layers. Alignment is reinforced through contrastive learning, where positive pairs consist of different modality representations of the same underlying content, while negative pairs come from different samples. This approach avoids modality-specific encoders or handcrafted fusion mechanisms, and instead allows the model to implicitly align modalities through feature-level supervision.

$\mathcal{L}_{recon}$ is the loss for reconstructing masked images, which is defined as:

$$\mathcal{L}_{recon} = \frac{1}{N}\sum_{i=1}^{N}|\widehat{x}_i - x_i|_2^2 \quad (3)$$

Where $\widehat{x}$ and x are the reconstructed and original images, respectively.

The final loss function for training our model is the combination of the three losses:

$$\mathcal{L} = \lambda_{\mathrm{img-text}}\mathcal{L}_{img-text} + \lambda_{\mathrm{img-img}}\mathcal{L}_{img-img} + \lambda_{\mathrm{recon}}\mathcal{L}_{recon} \quad (4)$$

Among them, $\lambda_{img-text}$ and $\lambda_{img-img}$ are set to 0.75, and λ_{recon} is set to 1, based on hyperparameter tuning experiments.

In EyeCLIP, all images share the same encoder, ensuring consistent feature extraction across different modalities. This innovative combination of CLIP and MAE distinguishes our approach from traditional CLIP models, enhancing its capability by utilizing a large amount of unlabeled data.

During the training phase of EyeCLIP, we cropped the images to field-of-view and resized them to 224×224, and applied data augmentation, including random resized cropping, color jitter, and horizontal flipping. The EyeCLIP was trained with a base learning rate of 0.001 for the first 2000 steps, with a 2-epoch warm-up, followed by cosine decay to zero throughout the training process. A batch size of 200 was used, and training was conducted on one NVIDIA Tesla V100 (32 GB) GPU for approximately four weeks. At the end of the training, the model with the lowest loss on the validation set was selected for testing.

Details of the Comparison Models

PubMedCLIP is a CLIP model specifically finetuned for the medical domain[25]. Trained on the Radiology Objects in COntext (ROCO) dataset[52], it encompasses over 80,000 samples from various medical imaging modalities like ultrasound, X-rays, computed tomography, magnetic resonance imaging, and various body regions. The texts used for training were the relatively short captions associated with the images in the ROCO dataset. Experimental outcomes showcased that leveraging PubMedCLIP as a pretrained visual encoder led to a potential performance boost of up to 3% for existing MedVQA models.

BioMedCLIP is a multimodal biomedical foundation model pretrained using 15 million scientific image-text pairs extracted from 4.4 million articles in PubMed Central[24]. It incorporates a domain-specific language model (PubMedBERT)[53], utilizes larger vision transformers, and integrates other domain-specific optimizations. Compared to general-domain CLIP and previous biomedical vision-language models such as PubMedCLIP, BioMedCLIP demonstrates superior performance across various downstream tasks, including cross-modal retrieval, zero-shot image classification, and VQA.

RETFound is trained on a vast dataset comprising 1.6 million unlabeled retinal images through self-supervised reconstruction[10]. It leveraged two ophthalmic modalities, CFP and OCT, to train separate weights for each modality. RETFound surpassed other comparative models, including those pretrained on ImageNet, in diagnosing sight-threatening eye conditions and predicting systemic disorders.

FLAIR[54] It is a pretrained vision-language model (ResNet50-based) for universal retinal fundus image understanding. It was trained using 37 open-access, mostly categorical fundus imaging datasets from various sources, with up to 97 different target conditions and 284, 660 images. It uses a textual expert's knowledge to describe the fine-grained features of the pathologies as well as the hierarchies and dependencies between them. It has been extensively validated to outperform more generalist, larger-scale image-language models such as CLIP or BiomedCLIP.

Downstream Validation Details

Zero-shot Classification. For zero-shot transfer, we followed the method in the CLIP experiment. Each class was associated with a text prompt consisting of the modality and class name (for example, 'color fundus, diabetic retinopathy'). We computed the ℓ2-normalized embedding using the text encoder and image encoder from EyeCLIP for the prompt and image. For each image, we computed the ℓ2-normalized embedding and then computed cosine-similarity scores between the image and each text embedding, and the predicted class was consequently the class with the highest similarity score.

Few-shot Classification. For few-shot classification, we varied the number of labeled examples per class for finetuning EyeCLIP (known as 'shot') from n = 1, 2, 4, 8, 16, and tested the model on the test set similar to full-data finetune classification.

Full-data Fine-tune Classification. We used each image encoder to extract a low-dimensional feature embedding from each image and added a multilayer perceptron to map the image feature representation to logits, which were interpreted as class probabilities after softmax normalization. During finetuning, the encoder was frozen for the first five epochs and unfrozen afterward. A total of 50 epochs was trained for each model. For single-label classification tasks, we used a batch size of 16. The first ten epochs implemented a learning rate warm-up from 0 to 5×10^{-4}, followed by a cosine annealing schedule reducing the learning rate from 5×10^{-4} to 1×10^{-6} over the remaining 40 epochs. Additionally, we adopted label smoothing cross-entropy loss with a smoothing factor of 0.1 to prevent the model from becoming overly confident in

dominant classes. For multi-label classification tasks in AngioReport and Retina Image Bank, we used a batch size of 4, trained for 30 epochs, and set the learning rate to 0.01. After each epoch, we evaluated the model on the validation set, saving the model weights with the highest AUROC for internal and external assessments.

Cross-Modal Retrieval. For cross-modal retrieval, we used the same method as zero-shot classification above to retrieve the top-K images that were closest in the aligned latent space to a specific text query (text-to-image retrieval). Image-to-text and image-to-image retrieval were performed analogously. To evaluate retrieval, we used Recall@K, which measures the percentage of correct results included in these top-K retrieved samples. We chose $K \in (1, 5, 10)$ and reported mean recall by averaging the scores over the three Recall@K values.

Visual Question Answering. For visual question answering, we used the image encoder from EyeCLIP to extract image features, which were then concatenated with text features (questions). The combined feature was fed into the language model Vicuna (Llama 2-7b)[55] for language generation, performing VQA.

To enhance multi-disease alignment, we employed the OpthalVQA (train) dataset for few-shot finetuning. Support examples per modality and disease were set to 1, 2, 4, and 8. For each scenario, five independent trials were performed with different random seeds to ensure robustness and reduce the influence of random variations in training set selection. We utilized the Low-Rank Adaptation (LoRA)[56] method for efficient finetuning, running for three epochs with an initial learning rate 2e-5. Cosine annealing was applied to adjust the learning rate dynamically. The model's performance was evaluated on the OpthalVQA (test) dataset using the checkpoint from the final epoch.

Evaluation Metrics

We employed the AUROC and AUPR metrics to assess the performance of classification tasks. These metrics gauge the classification effectiveness based on the receiver operating characteristics and precision-recall curves. When dealing with binary classification tasks, such as ocular disease diagnosis, we computed AUROC and AUPR in a binary context. For multi-class classification tasks such as five-stage DR and multi-class disease diagnosis, we calculated AUROC and AUPR individually for each disease class and then averaged them (macro) to derive the overall AUROC and AUPR scores.

Regarding VQA tasks, we utilized various classification-based metrics to evaluate performance, including the exact match score, F1 score, precision, recall, and language-based metric metrics such as BLEU[57] and sentence similarity.

For retrieval tasks, we used the metric Recall@K, which is the proportion of the data correctly retrieved among the top-K retrieved samples.

Statistical Analysis

We employed descriptive statistical methods to analyze demographic data, including age and gender. Two-sided t-tests were used to compare the AUROC and AUPR of EyeCLIP with those of other models (CLIP, BioMedCLIP, PubMedCLIP, RETFound, or FLAIR), selecting the most competitive model in each task based on mean performance to determine statistical significance. To enhance the robustness of the results, we conducted multiple trials using five different random seeds in the classification and VQA tasks. The final results were reported as the mean of these five trials, with the 95% confidence interval (CI) calculated using 1.96 × standard error.

Data availability

We do not have permission to redistribute the datasets used for developing EyeCLIP, the data may be available under constrained access from the corresponding author upon reasonable request. Downstream datasets can be accessed via the links: IDRID (https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid), MESSIDOR2 (https://www.adcis.net/en/third-party/messidor2/), APTOS-2019 (https://www.kaggle.com/c/aptos2019-blindness-detection/overview), PAPILA (https://figshare.com/articles/dataset/PAPILA/14798004/1), Glaucoma Fundus (https://doi.org/10.7910/DVN/1YRRAC), JSIEC (https://zenodo.org/record/3477553), Retina (https://www.kaggle.com/datasets/jr2ngb/cataractdataset), OCTID (https://borealisdata.ca/dataverse/OCTID), OCTDL (https://ieee-dataport.org/documents/octdl-optical-coherence-tomography-dataset-image-based-deep-learning-methods), AngioReport (https://tianchi.aliyun.com/dataset/170128), Retina Image Bank (https://imagebank.asrs.org/), OphthalVQA (https://figshare.com/s/3e8ad50db900e82d3b47).

Code availability

Code available at https://github.com/Michi-3000/EyeCLIP.

Received: 17 January 2025; Accepted: 3 June 2025;
Published online: 21 June 2025

References

1. Burton, M. J. et al. The Lancet Global Health Commission on Global Eye Health: vision beyond 2020. *Lancet Glob. Health* **9**, e489–e551 (2021).
2. Resnikoff, S. et al. Estimated number of ophthalmologists worldwide (International Council of Ophthalmology update): will we meet the needs? *Br. J. Ophthalmol.* **104**, 588–592 (2020).
3. Ye, J., He, L. & Beestrum, M. Implications for implementation and adoption of telehealth in developing countries: a systematic review of China's practices and experiences. *NPJ Digital Med.* **6**, 174 (2023).
4. Cui, M. & Zhang, D. Y. Artificial intelligence and computational pathology. *Lab. Investig.* **101**, 412–422 (2021).
5. Vorontsov, E. et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat. Med* **30**, 2924–2935 (2024).
6. Li, J.-P. O. et al. Digital technology, tele-medicine and artificial intelligence in ophthalmology: a global perspective. *Prog. Retinal Eye Res.* **82**, 100900 (2021).
7. Ting, D. S. W. et al. Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images From Multiethnic Populations With Diabetes. *JAMA* **318**, 2211–2223 (2017).
8. Chen, X. et al. FFA-GPT: an automated pipeline for fundus fluorescein angiography interpretation and question-answer. *npj Digital Med.* **7**, 111 (2024).
9. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
10. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
11. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual–language foundation model for pathology image analysis using medical Twitter. *Nat. Med* **29**, 2307–2316 (2023).
12. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med* **30**, 850–862 (2024).
13. Chia, M. A. et al. Foundation models in ophthalmology. *British Journal of Ophthalmology* (2024).
14. Yang, J. Multi-task learning for medical foundation models. *Nat. Comput Sci.* **4**, 473–474 (2024).
15. Qiu, J. et al. Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence. *NEJM AI* **1**, AIoa2300221 (2024).
16. Shi, D. et al. EyeFound: A Multimodal Generalist Foundation Model for Ophthalmic Imaging. *arXiv preprint arXiv:2405.11338* (2024).
17. Nath, S., Marie, A., Ellershaw, S., Korot, E. & Keane, P. A. New meaning for NLP: the trials and tribulations of natural language processing with GPT-3 in ophthalmology. *Br. J. Ophthalmol.* **106**, 889–892 (2022).
18. Shi, D. et al. Translation of Color Fundus Photography into Fluorescein Angiography Using Deep Learning for Enhanced Diabetic Retinopathy Screening. *Ophthalmol. Sci.* **3**, 100401 (2023).

19. Song, F., Zhang, W., Zheng, Y., Shi, D. & He, M. A deep learning model for generating fundus autofluorescence images from color fundus photography. *Adv. Ophthalmol. Pr. Res* **3**, 192–198 (2023).
20. Chen, R. et al. Translating color fundus photography to indocyanine green angiography using deep-learning for age-related macular degeneration screening. *npj Digital Med.* **7**, 34 (2024).
21. Shi, D. et al. Cross-modality Labeling Enables Noninvasive Capillary Quantification as a Sensitive Biomarker for Assessing Cardiovascular Risk. *Ophthalmol. Sci.* **4**, 100441 (2024).
22. Chen, X. et al. Evaluating large language models and agents in healthcare: key challenges in clinical applications. *Intelligent Medicine*, (2025).
23. Radford, A. et al. Learning transferable visual models from natural language supervision. in *International conference on machine learning* 8748-8763 (PMLR, 2021).
24. Zhang, S. et al. A Multimodal Biomedical Foundation Model Trained from Fifteen Million Image–Text Pairs. *NEJM AI* 2, https://doi.org/10.1056/AIoa2400640 (2024).
25. Eslami, S., Meinel, C. & De Melo, G. Pubmedclip: How much does clip benefit visual question answering in the medical domain? in *Findings of the Association for Computational Linguistics: EACL 2023* 1181-1193 (2023).
26. Gholami, P., Roy, P., Parthasarathy, M. K. & Lakshminarayanan, V. OCTID: Optical coherence tomography image database. *Computers Electr. Eng.* **81**, 106532 (2020).
27. Kulyabin, M. et al. OCTDL: Optical Coherence Tomography Dataset for Image-Based Deep Learning Methods. *Sci. Data* **11**, 365 (2024).
28. Zhang, W. et al. Angiographic Report Generation for the 3rd APTOS's Competition: Dataset and Baseline Methods. *medRxiv*, 2023-2011 (2023).
29. Retina Image Bank, available at https://imagebank.asrs.org/.
30. Gupta, K. & Reddy, S. Heart, Eye, and Artificial Intelligence: A Review. *Cardiol. Res* **12**, 132–139 (2021).
31. Yusufu, M. et al. Retinal vascular fingerprints predict incident stroke: findings from the UK Biobank cohort study. *Heart*, heartjnl-2024-324705 (2025).
32. Chua, S. Y. L. et al. Cohort profile: design and methods in the eye and vision consortium of UK Biobank. *BMJ Open* **9**, e025077 (2019).
33. Lu, M. Y. et al. A visual-language foundation model for computational pathology. *Nat. Med.* **30**, 863–874 (2024).
34. Xu, P., Chen, X., Zhao, Z. & Shi, D. Unveiling the clinical incapabilities: a benchmarking study of GPT-4V(ision) for ophthalmic multimodal image analysis. *British Journal of Ophthalmology*, bjo-2023-325054 (2024).
35. Savage, T., Nayak, A., Gallo, R., Rangan, E. & Chen, J. H. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Med.* **7**, 20 (2024).
36. Chen, X. et al. ICGA-GPT: report generation and question answering for indocyanine green angiography images. *Br. J. Ophthalmol.* **108**, 1450–1456 (2024).
37. Chen, X. et al. EyeGPT for Patient Inquiries and Medical Education: Development and Validation of an Ophthalmology Large Language Model. *J. Med. Internet Res.* **26**, e60063 (2024).
38. Zhao, Z. et al. Slit Lamp Report Generation and Question Answering: Development and Validation of a Multimodal Transformer Model with Large Language Model Integration. *J. Med Internet Res* **26**, e54047 (2024).
39. Gende, M. et al. Automatic Segmentation of Retinal Layers in Multiple Neurodegenerative Disorder Scenarios. *IEEE J. Biomed. Health Inform.* **27**, 5483–5494 (2023).
40. Shi, D. et al. A Deep Learning System for Fully Automated Retinal Vessel Measurement in High Throughput Image Analysis. *Front Cardiovasc Med* **9**, 823436 (2022).
41. Li, C. et al. Retinal oculomics and risk of incident aortic aneurysm and aortic adverse events: a population-based cohort study. *Int J Surg*, (2025).
42. Wu, Y. et al. Noninvasive early prediction of preeclampsia in pregnancy using retinal vascular features. *npj Digital Med.* **8**, 188 (2025).
43. Zhang, W. et al. Fundus2Video: Cross-Modal Angiography Video Generation from Static Fundus Photography with Clinical Knowledge Guidance. in *Medical Image Computing and Computer Assisted Intervention – MICCAI* 689-699 (Springer Nature Switzerland, Morocco, 2024).
44. Wu, X. et al. FFA Sora, video generation as fundus fluorescein angiography simulator. *arXiv preprint arXiv:2412.17346* (2024).
45. Polino, A., Pascanu, R. & Alistarh, D.-A. Model compression via distillation and quantization. *in 6th International Conference on Learning Representations*, (2018).
46. Shi, D., He, S., Yang, J., Zheng, Y. & He, M. One-shot Retinal Artery and Vein Segmentation via Cross-modality Pretraining. *Ophthalmol. Sci.* **4**, 100363 (2024).
47. Porwal, P. et al. Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. *Data* **3**, 25 (2018).
48. Kovalyk, O. et al. PAPILA: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. *Sci. Data* **9**, 291 (2022).
49. Ahn, J. M. et al. A deep learning model for the detection of both advanced and early glaucoma using fundus photography. *PloS one* **13**, e0207982 (2018).
50. 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).
51. He, K. et al. Masked autoencoders are scalable vision learners. 16000-16009.
52. Pelka, O., Koitka, S., Rückert, J., Nensa, F. & Friedrich, C. M. Radiology objects in context (roco): a multimodal image dataset. in *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3* 180-189 (Springer, 2018).
53. Gu, Y. et al. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc. (HEALTH)* **3**, 1–23 (2021).
54. Silva-Rodriguez, J., Chakor, H., Kobbi, R., Dolz, J. & Ayed, I. B. A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. *Med. Image Anal.* **99**, 103357 (2025).
55. Touvron, H. et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. (2023).
56. Hu, E. J. et al. LoRA: Low-Rank Adaptation of Large Language Models. in *International Conference on Learning Representations* (2021).
57. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation. in *the 40th Annual Meeting* 311 (Association for Computational Linguistics, 2001).

Acknowledgements

We thank the American Society of Retina Specialists for providing the valuable Retina Image Bank and InnoHK HKSAR Government for valuable supports. The research described in this paper was conducted in the JC STEM Lab of Innovative Light Therapy for Eye Diseases funded by The Hong Kong Jockey Club Charities Trust. The study was supported by the Global STEM Professorship Scheme (P0046113) and Henry G. Leong Endowed Professorship in Elderly Vision Health (PI: Mingguang He). The sponsor or funding organization had no role in the design or conduct of this research.

Author contributions

D.S., J.Y., S.H., and M.H. conceived the study. D.S. and W.Z. built the deep learning model and ran experiments. M.H. provided data and computing facilities. D.S., W.Z., J.Y., S.H., and X.C. contributed to key data interpretation. D.S. W.Z, X.C. wrote the manuscript. P.X., K.J., S.L., J.W., M.Y., S.L., Q.Z., Z.G., X.X., and M.H. critically revised the manuscript. All authors have read and approved the manuscript.

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/s41746-025-01772-2.

Correspondence and requests for materials should be addressed to Danli Shi or Mingguang He.

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