npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-026-02560-2

Boosting foundation models for rare eye disease diagnosis via a multimodal text-to-image generative framework

Check for updates

Ruoyu Chen[1,4], Weiyi Zhang[1,4], Bowen Liu[1,4], Xinyuan Wu[1], Xiaolan Chen[1], Pusheng Xu[1], Shunming Liu[1], Mingguang He[1,2,3] ✉ & Danli Shi[1,2] ✉

The rising prevalence of vision-threatening retinal diseases poses a significant burden on the global healthcare systems. Though deep learning (DL) techniques offer promising avenues for improving diagnostic efficiency, data scarcity and imbalance issues persist in training robust diagnostic models, particularly for rare eye diseases. Here, we introduce EyeDiff, a generative foundation model capable of synthesizing lesion-preserving ophthalmic images from textual descriptions. Both objective metrics and expert human evaluations confirmed EyeDiff's ability to generate high-fidelity images across multiple imaging modalities, accurately reflecting textual descriptions of diverse retinal diseases and lesion types. By augmenting minority classes across 11 globally sourced datasets, EyeDiff consistently boosted the diagnostic accuracy for both common and rare eye diseases across different foundation model types, including modality-specific, multimodal and vision-language foundation models trained solely on real data. These results underscore EyeDiff's potential as a general-purpose text-to-image foundation model, offering a scalable and flexible approach to generate balanced, disease-relevant data for advancing retinal disease diagnosis.

The increasing prevalence of vision-threatening retinal diseases has significantly strained the global healthcare system[1]. Multimodal ophthalmic images, such as color fundus photographs (CFP), optical coherence tomography (OCT), fundus fluorescein angiography (FFA), etc. provide complementary information on ocular abnormalities[2–4]. While deep learning (DL) techniques have shown great promise in enhancing diagnostic accuracy and efficiency[5,6], growing concerns over data privacy and institutional restrictions on data sharing make it increasingly difficult to integrate datasets across clinical centers and further exacerbating the problem of data imbalance in training robust diagnostic models[7–9]. Recent foundation models have shown remarkable diagnostic performance with few-shot fine-tuning, marking a significant advancement in data efficiency. However, their accuracy for minority and rare classes remains suboptimal because pretraining on extremely limited and imbalanced data prevents the models from effectively learning specific disease features.

To mitigate data scarcity and imbalance, various data augmentation techniques have been explored[10,11]. Traditional methods, such as random oversampling, address class imbalance by replicating minority samples[12], but this can lead to overfitting and reduced generalization[13]. More recently, generative adversarial networks (GANs) have been widely employed to synthesize realistic fundus images, with the aim of enriching existing datasets and enhancing DL-based diagnostic performance[14–19]. Although effective, most GAN-based approaches are limited to unimodal image generation and tailored for single-task classification, restricting their broader utility. There remains a pressing need for generalist generative models capable of addressing the diverse data needs of ophthalmic applications.

Stable Diffusion (SD), a recent advancement in text-to-image generation, offers a powerful solution by producing high-quality, diverse images from natural language prompts[20]. In general computer vision, diffusion-generated images have been shown to rival or surpass real images for self-supervised model training[21]. This generative paradigm holds significant promise for building universal models capable of synthesizing large-scale, richly annotated datasets from textual input alone, effectively generating digital twin datasets[22,23]. Despite this potential, its application in ophthalmology, particularly for producing multimodal diagnostic images, remains underexplored.

Here, we present EyeDiff, a multimodal text-to-image generative foundation model for synthesizing ophthalmic images from descriptive text prompts. Trained on paired textual and image data across 14 imaging

[1]School of Optometry, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China. [2]Research Centre for SHARP Vision, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China. [3]Centre for Eye and Vision Research (CEVR), 17W Hong Kong Science Park, Hong Kong SAR, China. [4]These authors contributed equally: Ruoyu Chen, Weiyi Zhang, Bowen Liu. ✉e-mail: mingguang.he@polyu.edu.hk; danli.shi@polyu.edu.hk

modalities, including CFP, FFA, OCT, indocyanine green angiography (ICGA), fundus autofluorescence (FAF), EyeDiff enables lesion-preserving, modality-aware generation of diagnostic-quality ophthalmic images. We further validate EyeDiff's generalizability by evaluating its performance in boosting different foundation models across 11 diverse global datasets for minority class augmentation and multi-disease diagnosis, covering both common and rare retinal diseases.

Results

A total of 42,048 images from 8 datasets were used for model development. The quality of EyeDiff-generated images was evaluated using quantitative metrics for text-image alignment and human assessments for visual quality and authenticity. To validate the model's generalizability, we used different established foundation models as baselines: (1) RETFound[24]: a modality-specific foundation model trained with separate weights for different imaging modalities using the Vision Transformer (ViT) architecture; (2) EyeFound[25]: a versatile foundation model trained with the ViT architecture, designed to learn generalizable representations from unlabeled multimodal retinal images; (3) EyeCLIP[26]: a multimodal visual-language foundation model trained on millions of multimodal ophthalmic images paired with partial clinical text using the contrastive language-image pre-training (CLIP) framework. We then evaluated whether augmenting minority and rare classes in downstream datasets with EyeDiff-generated images could improve the classification performance of these foundation models for both common and rare diseases. A total of 4403 images from 2 internal datasets and 13,198 images from 9 external datasets were enrolled for downstream tasks. Detailed characteristics of the datasets and the flowchart of the study are shown in Table 1 and Fig. 1.

Quantitative evaluation

We curated 77 text-image category pairs from the downstream datasets to objectively evaluate the quality of EyeDiff-generated images. The Visual Question Answering Score (VQAScore)[27] was used to assess the alignment between the generated images and their corresponding textual descriptions. As shown in Table 2, the average VQAScore was 0.822 for OCT-based disease detection, 0.776 for CFP-based multi-category eye disease diagnosis, and 0.670 for multimodal imaging-based rare disease diagnosis. These results demonstrate a strong correspondence between the generated images and textual prompts across various imaging modalities and diagnostic tasks.

Medical realism evaluation

The results of the Turing test and visual quality scores provided by two ophthalmologists are as follows.

Turing test

To evaluate the medical realism of EyeDiff-generated images, a Turing test was conducted using 100 images (50 real and 50 EyeDiff-generated). Grader 1 achieved an overall correct classification rate of 50.00%, while grader 2 achieved 54.50%. For real images, the identification accuracies

Table 1 | Characteristics of datasets used for model development and validation

Dataset	Image Modality	Disease Category	*N* (%)
Generative Model Development			42,048 (100.00%)
Retina Image Bank	CFP, FFA, ICGA, OCT, FAF, RetCam, B scan, Slit Lamp, SLO, WF-SLO, UWF-SLO, External eye, Red free, OUS	84 ocular diseases	22,941 (54.56%)
EyePACS	CFP	DR	15,202 (36.15%)
OCTDL	OCT	AMD, DME, RVO, RAO, ERM, VID, Normal	2064 (4.91%)
REFUGE	CFP	Glaucoma	1200 (2.85%)
ORIGA	CFP	Glaucoma	282 (0.67%)
RIM-ONE	CFP	Glaucoma, Normal	158 (0.38%)
DRISHTI	CFP	Glaucoma, Normal	101 (0.24%)
GAMMA	CFP, OCT	Glaucoma	100 (0.24%)
Downstream Validation			
Internal Validation			4403 (100.00%)
Retina Image Bank (Rare Diseases)	CFP, FFA, ICGA, OCT, FAF, RetCam, B scan, Slit Lamp, SLO, WF-SLO, UWF-SLO, External eye, Red free	17 rare diseases	2339 (53.12%)
OCTDL	OCT	AMD, DME, RVO, RAO, ERM, VID, Normal	2064 (46.88%)
External Validation			13,198 (100.00%)
APTOS2019	CFP	DR	3662 (27.75%)
OphthalWeChat	CFP, FFA, ICGA, OCT, FAF, RetCam, B scan, Slit Lamp, SLO, WF-SLO, UWF-SLO, External eye, Red free	178 common and rare diseases	3071 (23.27%)
MESSIDOR-2	CFP	DR	1744 (13.21%)
Glaucoma Fundus	CFP	Glaucoma	1544 (11.70%)
JSIEC	CFP	39 fundus diseases	1000 (7.58%)
Retina	CFP	Glaucoma, Cataract, Normal	601 (4.55%)
OCTID	OCT	MH, CSCR, AMD DR, Normal	572 (4.33%)
IDRiD	CFP	DR	516 (3.91%)
PAPILA	CFP	Glaucoma	488 (3.70%)

OCT Optical coherence tomography, *CFP* Color fundus photography, *FFA* Fundus fluorescein angiography, *ICGA* Indocyanine green angiography, *FAF* Fundus autofluorescence, *SLO* Scanning laser ophthalmoscopy, *WF-SLO* Widefield SLO, *UWF-SLO* Ultra-widefield SLO, *OUS* Ocular ultrasound, *DR* Diabetic retinopathy, *AMD* Age-related macular degeneration, *DME* Diabetic macular edema, *ERM* Epiretinal Membrane, *RVO* Retinal vein occlusion, *RAO* Retinal artery occlusion, *VID* Vitreoretinal interface disease, *MH* Macular hole, *CSCR* Central serous chorioretinopathy.

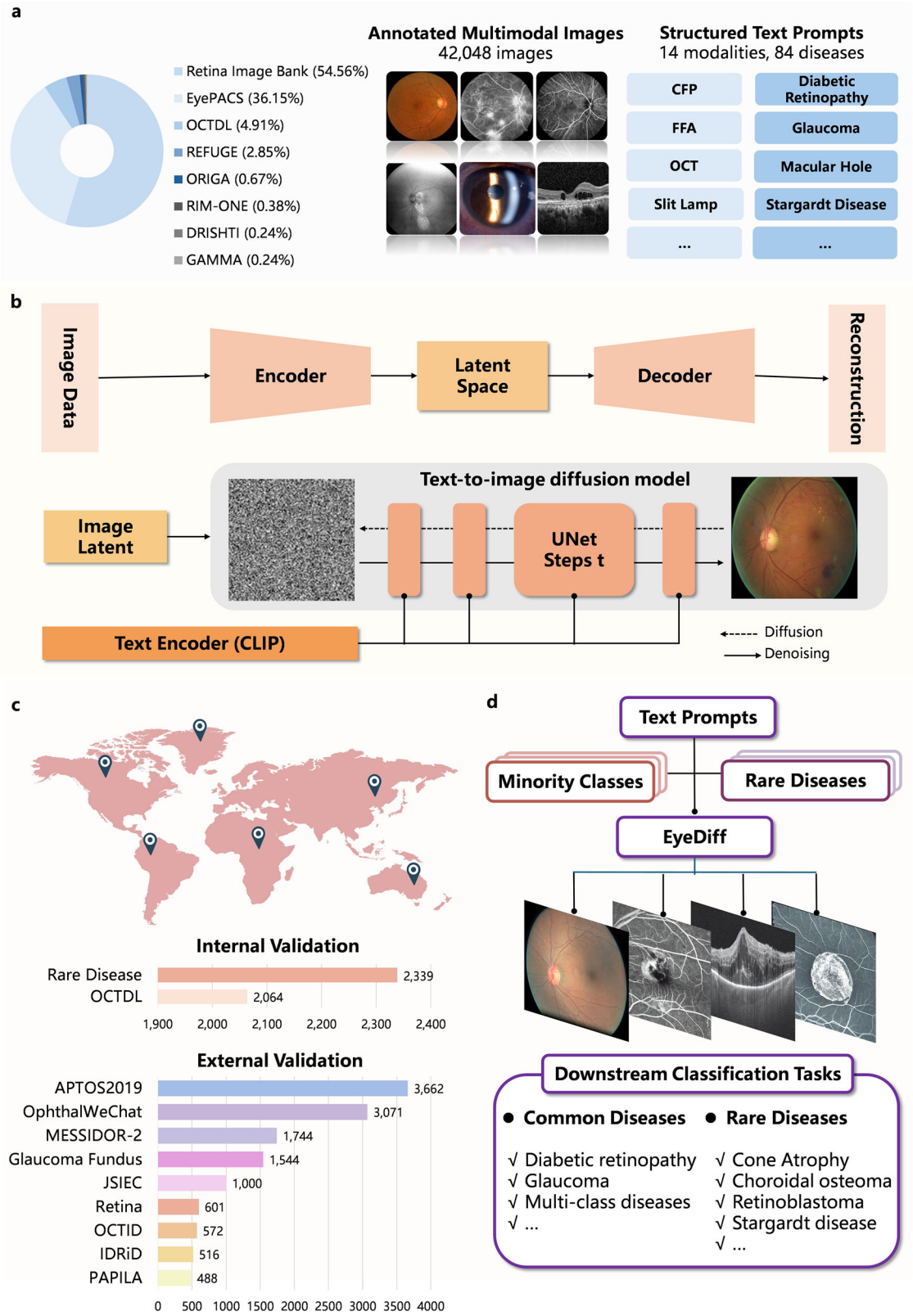

Fig. 1 | Flow diagram of this study. a Characteristics of training datasets. **b** The architecture of the text-to-image diffusion model. Given a retinal image, the encoder transforms it into a latent representation, while the decoder reconstructs the image from this latent. The text-to-image diffusion model architecture is formed by a latent diffusion model (LDM) that consists of time-conditional UNets. The LDM is conditioned on text embeddings derived from the CLIP model. During the training phase, noise is progressively applied to the image latent according to a specified noise scheduler at each timestep *t*. The LDM learns to reduce this noise given the noisy image latent, the timestep t and the text embedding as inputs. In the inference phase, it begins with completely random noise, guided by a sampling timestep value and a text embedding. Finally, the decoder reconstructs a new synthetic image from the denoised image latent. **c** Downstream dataset characteristics. **d** The proportion of internal and external validation datasets for downstream disease classifications. CFP color fundus photography, OCT optical coherence tomography, FFA fundus fluorescein angiography.

Table 2 | VQAScore results for assessing text-image alignment of EyeDiff-generated images

Target dataset	Prompts used in downstream tasks	VQAScore
OCT-based disease detection	optical coherence tomography, retinal artery occlusion	0.764
	optical coherence tomography, macular hole	0.789
	optical coherence tomography, vitreomacular interface disease	0.850
	optical coherence tomography, diabetic macular edema	0.827
	optical coherence tomography, diabetic retinopathy	0.762
	optical coherence tomography, epiretinal membrane	0.888
	optical coherence tomography, central serous chorioretinopathy	0.841
	optical coherence tomography, normal	0.794
	optical coherence tomography, retinal vein occlusion	0.850
	optical coherence tomography, age-related macular degeneration	0.857
	[a]Average	0.822
CFP-based multi-category eye disease classification	fundus image, tessellated fundus	0.861
	fundus image, cataract	0.824
	fundus image, moderate non-proliferative diabetic retinopathy	0.841
	fundus image, blur fundus with suspected proliferative diabetic retinopathy	0.856
	fundus image, retinal disease	0.842
	fundus image, retinal artery occlusion	0.636
	fundus image, disc swelling and elevation	0.795
	fundus image, peripheral retinal degeneration and break	0.819
	fundus image, blur fundus	0.805
	fundus image, vessel tortuosity	0.690
	fundus image, macular hole	0.830
	fundus image, large optic cup	0.741
	fundus image, branch retinal vein occlusion	0.747
	fundus image, non-referable diabetic retinopathy	0.721
	fundus image, possible glaucoma	0.882
	fundus image, suspected glaucoma	0.845
	fundus image, congenital disc abnormality	0.684
	fundus image, yellow-white spots-flecks	0.834
	fundus image, severe and proliferative diabetic retinopathy	0.863
	fundus image, bietti crystalline dystrophy	0.804
	fundus image, epiretinal membrane	0.776
	fundus image, vitreous particles	0.708
	fundus image, central serous chorioretinopathy	0.844
	fundus image, laser spots	0.791
	fundus image, glaucoma	0.827
	fundus image, vogt-koyanagi-harada disease	0.829
	fundus image, myelinated nerve fiber	0.717
	fundus image, central retinal vein occlusion	0.662
	fundus image, optic atrophy	0.704
	fundus image, silicon oil in eye	0.816
	fundus image, advanced glaucoma	0.837
	fundus image, rhegmatogenous retinal detachment	0.628
	fundus image, maculopathy	0.767
	fundus image, coloboma	0.796
	fundus image, massive hard exudates	0.772
	fundus image, severe hypertensive retinopathy	0.826
	fundus image, pathological myopia	0.874
	fundus image, fundus neoplasm	0.660
	fundus image, cotton-wool spots	0.829
	fundus image, preretinal hemorrhage	0.720
	fundus image, early glaucoma	0.801
	fundus image, dragged disc	0.805
	fundus image, retinitis pigmentosa	0.768
	fundus image, fibrosis	0.795
	fundus image, mild diabetic retinopathy	0.713
	fundus image, severe diabetic retinopathy	0.889
	fundus image, normal	0.274
	fundus image, moderate diabetic retinopathy	0.819
	fundus image, proliferative diabetic retinopathy	0.864
	[a]Average	0.776
Multimodal imaging-based rare disease classification	birdshot retinochoroidopathy	0.675
	pseudoxanthoma elasticum	0.629
	familial exudative vitreoretinopathy	0.683
	macular telangiectasia	0.751
	central areolar choroidal dystrophy	0.788
	optic disc pit	0.761
	retinoblastoma	0.589
	Stargardt disease	0.724
	Stargardt disease, cone dystrophy	0.694
	acute posterior multifocal placoid pigment epitheliopathy	0.793
	serpiginous choroiditis	0.648

Table 2 (continued) | VQAScore results for assessing text-image alignment of EyeDiff-generated images

Target dataset	Prompts used in downstream tasks	VQAScore
	serpiginous choroiditis, acute posterior multifocal placoid pigment epitheliopathy	0.704
	optic nerve hypoplasia	0.586
	choroidal osteoma	0.556
	retinopathy of prematurity	0.493
	congenital hypertrophy of the retinal pigment epithelium	0.688
	cone dystrophy	0.652
	choroidal melanoma	0.565
	retinitis pigmentosa	0.673
	retinitis pigmentosa, cone dystrophy	0.757
	[a]Average	0.670

OCT Optical coherence tomography, *CFP* Color fundus photography.
[a]Average indicates the average VQAScore of the corresponding datasets.

Table 3 | Turing test results for EyeDiff-generated images

Image modality	Number of images (*N*)	Number of realistic images (*N*, %)	
		Grader 1	Grader 2
CFP	15	9 (60.00%)	10 (66.67%)
OCT	16	10 (62.50%)	11 (68.75%)
FFA	15	9 (60.00%)	10 (66.67%)
FAF	2	2 (100.00%)	2 (100.00%)
UWF CFP	2	1 (50%.00)	0 (0.00%)
Total	50	31 (62.00%)	33 (66.00%)

Realistic images were defined as those generated by EyeDiff that were mistaken for real images by human graders in a blinded assessment.
CFP color fundus photography, *OCT* optical coherence tomography, *FFA* fundus fluorescein angiography, *FAF* fundus autofluorescence, *UWF CFP* ultrawide field color fundus photography.

were 62.00% (grader 1) and 76.00% (grader 2); for generated images, accuracies were 38.00% and 33.33%, respectively. Notably, 62.00% to 66.67% of generated images were mistaken for real ones. By modality, the proportion of EyeDiff images misidentified as real was 60.00% to 66.67% for CFP images, 62.50% to 68.75% for OCT images, 60.00% to 66.67% for FFA images, 100% for FAF images, and 50.00% to 100% for ultra-widefield color fundus photograph (UWF-CFP) (see Table 3). Most of the generated images demonstrated considerable authenticity. The common distinguishing features between the generated and real images included unrealistic colors, enhanced edges of retinal structures or lesions, and high levels of noise.

Visual quality evaluation

Fifty generated images were randomly selected for visual quality assessment by two experienced ophthalmologists (R.C. and X.C.) using a five-point scale (1 = generated images fully capture the key elements described in the text prompt; 5 = generated images do not contain any of the key elements described in the text prompt). The mean ± standard deviation visual quality scores were 1.940 ± 1.085 for grader 1 and 2.080 ± 1.055 for grader 2, with an inter-rater agreement of Kappa = 0.870, indicating high consistency. These results support the model's ability to synthesize lesion-preserving, modality-consistent images that align closely with clinical descriptions. Representative examples of text prompts and their corresponding EyeDiff-generated images are shown in Fig. 2.

Downstream performance for common vision-threatening diseases

Incorporating EyeDiff-generated images alongside original real images significantly enhanced diagnostic performance for both diabetic retinopathy (DR) and glaucoma when using RETFound as the baseline model.

For DR diagnosis using the IDRiD dataset[28], the area under the receiver operating curve (AUROC) improved from 0.826 (95% confidence interval [CI]: 0.821–0.832) at baseline to 0.837 (95% CI: 0.833–0.840) after augmentation with EyeDiff-generated images. Similarly, the area under the precision-recall curve (AUPR) rose from 0.502 (95% CI: 0.483–0.520) to 0.518 (95% CI: 0.509–0.527), outperforming models augmented by traditional oversampling techniques. For glaucoma diagnosis using the Glaucoma Fundus dataset[29], the AUROC was 0.950 (95% CI: 0.937–0.964) for original images, rising to 0.954 (95% CI: 0.940–0.968) with oversampling and 0.959 (95% CI: 0.945–0.973) after EyeDiff augmentation. The AUPR improved from 0.876 (95% CI: 0.841–0.911) to 0.879 (95% CI: 0.843–0.915) with oversample augmentation and 0.893 (95% CI: 0.855–0.931) with EyeDiff-generated images. These improvements were statistically significant (Table 4).

Performance improvements in DR and glaucoma diagnosis after augmentation with EyeDiff-generated images have also been observed using EyeFound and EyeCLIP as baseline models (Supplementary Table 1 and Supplementary Table 2).

Downstream performance for multi-class disease diagnosis

The integration of EyeDiff-generated images significantly enhanced model performance in multi-class disease detection tasks across various foundation models, including the modality-specific RETFound, multimodal EyeFound, and vision-language foundation model EyeCLIP.

For CFP-based diagnosis using RETFound (CFP weight) as baseline: On the JSIEC dataset[30], the AUROC increased from 0.990 (95% CI: 0.989–0.992) at baseline to 0.996 (95% CI: 0.995–0.997) after EyeDiff augmentation, while the AUPR rose from 0.887 (95% CI: 0.871–0.891) to 0.967 (95% CI: 0.957–0.978). On the Retina dataset, AUROC improved from 0.857 (95% CI: 0.831–0.873) at baseline to 0.892 (95% CI: 0.867–0.918) after EyeDiff augmentation, and AUPR increased from 0.720 (95% CI: 0.688–0.761) to 0.779 (95% CI: 0.731–0.826). All improvements were statistically significant.

For OCT-based diagnosis using RETFound (OCT weight) as baseline: On the OCTID dataset[31], AUROC increased from 0.993 (95% CI: 0.987–0.999) at baseline to 0.995 (95% CI: 0.992–0.997) after EyeDiff augmentation, and AUPR from 0.980 (95% CI: 0.967–0.993) to 0.982 (95% CI: 0.969–0.994). On the OCTDL dataset[32], AUROC improved from 0.982 (95% CI: 0.972–0.992) at baseline to 0.996 (95% CI: 0.995–0.997) after EyeDiff augmentation, and AUPR from 0.903 (95% CI: 0.862–0.925) to 0.967 (95% CI: 0.957–0.978).

Similarly, EyeDiff-generated images also boosted the performance of multimodal foundation models in multi-class disease detection (Supplementary Table 1 and Supplementary Table 2). These consistent performance gains across modalities and datasets underscore the robustness and generalizability of EyeDiff for augmenting diverse ophthalmic diagnostic tasks (see Table 4).

Downstream performance for under-represented disease subtype diagnosis

A minority class refers to a category of images that is significantly under-represented compared to other subtypes in a dataset, leading to data imbalance issues in downstream classification tasks. Data imbalance issues exist in the following datasets, and the specific minority classes are shown in Table 5. Supplementing EyeDiff-generated images significantly enhances disease diagnosis performance in these imbalanced classes. (1) IDRiD dataset ($n = 103$): The minority class comprised images with mild retinopathy ($n = 5$). The AUROC for mild retinopathy increased from 0.772 (95% CI: 0.733–0.811) at baseline

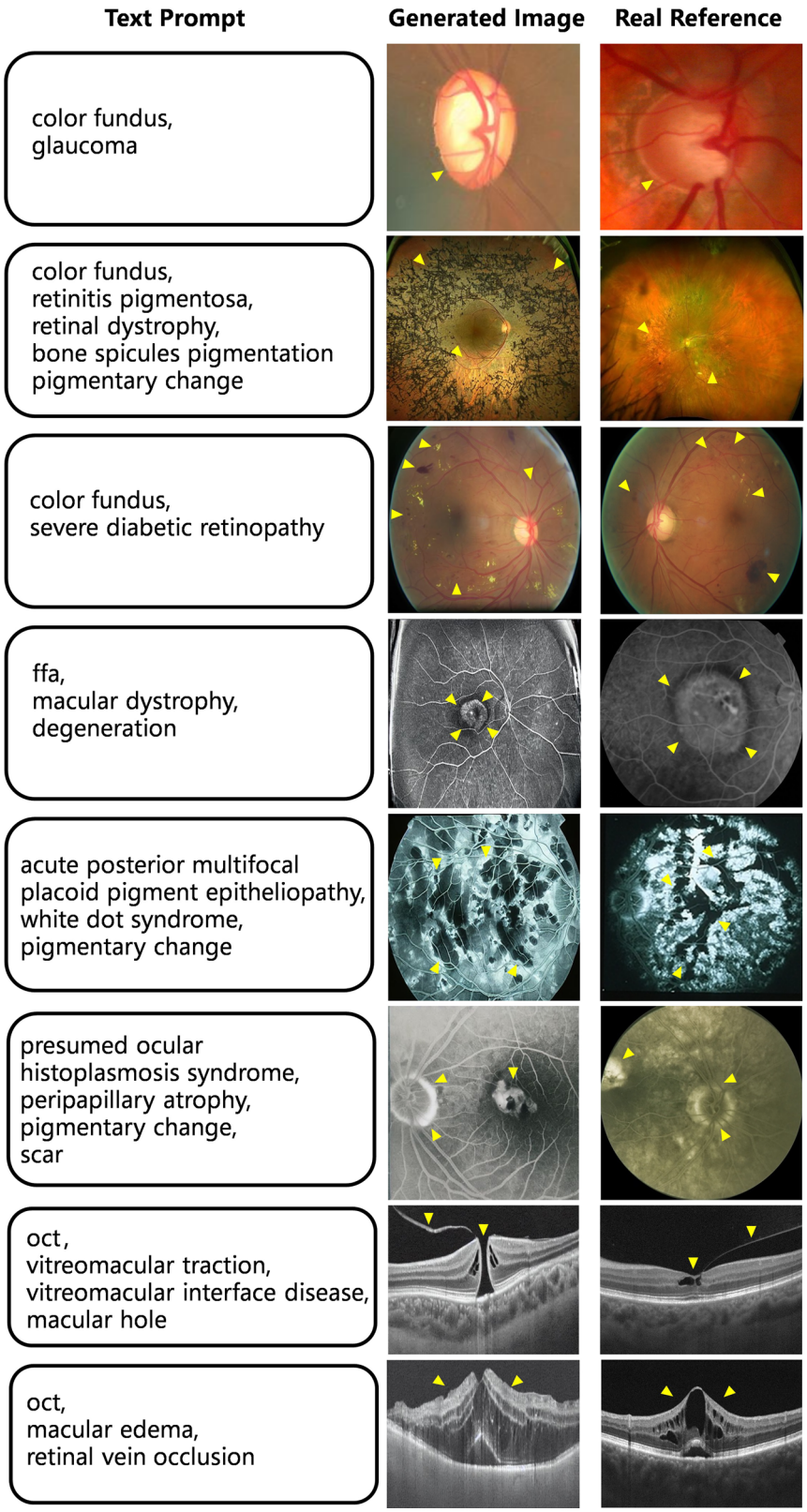

Fig. 2 | Examples of lesion-preserving generated images using text prompts. OCT optical coherence tomography; FFA fundus fluorescein angiography. The yellow arrow outlines lesions. 1st row: enlarged cup-disc ratio; 2nd row: bone spicules pigmentation and retinal dystrophy; 3rd row: microaneurysm, exudation, retinal hemorrhage, intraretinal microvascular abnormality; 4th row: macular dystrophy and degeneration; 5th row: characteristic lesion of acute posterior multifocal placoid pigment epitheliopathy; 6th row: macular lesion, peripapillary atrophy, pigmentary change; 7th row: vitreomacular traction and macular hole; 8th row: macular edema.

Table 4 | Overall performance comparison of the baseline RETFound model, oversampling, and EyeDiff-based synthetic augmentation in downstream classification tasks

Dataset	Model	AUROC (95% CI)	AUPR (95% CI)	*P* value
IDRiD	RETFound	0.826 (0.821, 0.832)	0.502 (0.483, 0.520)	
	Oversample	0.833 (0.826, 0.841)	0.516 (0.499, 0.532)	**0.012***
	EyeDiff	0.837 (0.833, 0.840) ↑	0.518 (0.509, 0.527) ↑	**0.008***
APTOS 2019	RETFound	0.946 (0.941, 0.950)	0.723 (0.703, 0.745)	
	Oversample	0.944 (0.941, 0.947)	0.709 (0.689, 0.729)	0.197
	EyeDiff	0.945 (0.941, 0.948)	0.714 (0.697, 0.732)	0.211
MESSIDOR2	RETFound	0.884 (0.880, 0.887)	0.669 (0.656, 0.683)	
	Oversample	0.856 (0.840, 0.872)	0.625 (0.609, 0.641)	0.187
	EyeDiff	0.879 (0.869, 0.890)	0.662 (0.631, 0.693)	0.213
Glaucoma Fundus	RETFound	0.950 (0.937, 0.964)	0.876 (0.841, 0.911)	
	Oversample	0.954 (0.940, 0.968)	0.879 (0.843, 0.915)	**0.009***
	EyeDiff	0.959 (0.945, 0.973) ↑	0.893 (0.855, 0.931) ↑	**0.008***
PAPILA	RETFound	0.820 (0.788, 0.854)	0.678 (0.646, 0.709)	
	Oversample	0.833 (0.806, 0.859)	0.724 (0.691, 0.757)	**<0.001***
	EyeDiff	0.814 (0.779, 0.848)	0.676 (0.638, 0.713)	**<0.001***
JSIEC	RETFound	0.990 (0.989, 0.992)	0.887 (0.871, 0.891)	
	Oversample	0.993 (0.990, 0.996)	0.937 (0.902, 0.951)	0.133
	EyeDiff	0.996 (0.995, 0.997) ↑	0.967 (0.957, 0.978) ↑	0.082
Retina	RETFound	0.857 (0.831, 0.873)	0.720 (0.688, 0.761)	
	Oversample	0.864 (0.852, 0.876)	0.731 (0.712, 0.750)	**<0.001***
	EyeDiff	0.892 (0.867, 0.918) ↑	0.779 (0.731, 0.826) ↑	**<0.001***
OCTID	RETFound	0.993 (0.987, 0.999)	0.980 (0.967, 0.993)	
	Oversample	0.993 (0.984, 1.000)	0.980 (0.958, 0.999)	**0.045***
	EyeDiff	0.995 (0.992, 0.997) ↑	0.982 (0.969, 0.994) ↑	**0.017***
OCTDL	RETFound	0.982 (0.972, 0.992)	0.903 (0.862, 0.925)	
	Oversample	0.994 (0.992, 0.996)	0.947 (0.928, 0.966)	**<0.001***
	EyeDiff	0.996 (0.995, 0.997) ↑	0.967 (0.957, 0.978) ↑	**<0.001***
ImageBank	RETFound	0.871 (0.863, 0.891)	0.439 (0.401, 0.462)	
	Oversample	0.893 (0.872, 0.923)	0.471(0.454, 0.501)	**<0.001***
	EyeDiff	0.919 (0.882, 0.931) ↑	0.530 (0.497, 0.550) ↑	**<0.001***
OphthalWeChat	RETFound	0.613 (0.570, 0.634)	0.397 (0.351, 0.431)	
	Oversample	0.650 (0.603, 0.692)	0.411(0.382, 0.437)	**<0.001***
	EyeDiff	0.663 (0.629, 0.701)↑	0.439 (0.401, 0.468)↑	**<0.001***

P values were compared to baseline RETFound.

CI confidence interval, *AUROC* Area under the receiver operating characteristic curve, *AUPR* Area under the precision-recall curve.

*$P < 0.05$.

using the original images to 0.817 (95% CI: 0.780–0.864) after incorporating EyeDiff-augmented images. (2) APTOS 2019 dataset ($n = 1100$): The minority class consisted of images with severe retinopathy ($n = 58$). The AUROC of severe retinopathy was increased from 0.867 (95% CI: 0.829–0.906) at baseline to 0.914 (95% CI: 0.903–0.934) after EyeDiff augmentation. (3) MESSIDOR-2 dataset ($n = 526$): The minority class consisted of images with proliferative retinopathy ($n = 11$). The AUROC for proliferative retinopathy increased from 0.960 (95% CI: 0.937–0.988) to 0.980 (95% CI: 0.967–0.994) after EyeDiff augmentation. (4) Glaucoma Fundus dataset ($n = 465$): The minority class comprised images with early glaucoma ($n = 87$). The AUROC of early glaucoma increased from 0.860 (95% CI: 0.827–0.873) to 0.927 (95% CI: 0.919–0.934) after EyeDiff augmentation. (5) PAPILA dataset ($n = 98$)[33]: The minority class comprised images with glaucoma ($n = 14$). The AUROC of glaucoma increased from 0.754 (95% CI: 0.742–0.778) to 0.795 (95% CI: 0.756–0.813) after EyeDiff augmentation.

Downstream performance for rare disease diagnosis

Adding EyeDiff-generated images into the Rare Diseases and OphthalWeChat[34] datasets significantly improved rare disease classification performance.

For the Rare Diseases dataset, the AUROC increased from 0.871 (95% CI: 0.863–0.891) at baseline to 0.893 (95% CI: 0.872–0.923) with oversampling and further to 0.919 (95% CI: 0.882–0.931) with EyeDiff augmentation. The AUPR improved from 0.439 (95% CI: 0.401–0.462) to 0.530 (95% CI: 0.497–0.550) with EyeDiff-generated images.

For the OphthaWeChat dataset, the AUROC rose from 0.613 (95% CI: 0.570–0.634) at baseline to 0.663 (95% CI: 0.629–0.701) with EyeDiff augmentation, and the AUPR from 0.397 (95% CI: 0.351–0.431) to 0.439 (95% CI: 0.401–0.468) (see Table 4).

Notably, the inclusion of EyeDiff-generated images significantly enhanced classification performance across multiple rare disease subgroups, including Stargardt disease, retinopathy of prematurity, retinoblastoma,

Table 5 | Comparison of baseline RETFound, oversampling, and EyeDiff-based synthetic augmentation in enhancing minority class performance in downstream classification tasks

Dataset	Minority Class	Model	AUROC (95% CI)	AUPR (95% CI)	*P* value
IDRiD (*n* = 103)	Mild Retinopathy (*n* = 5)	RETFound	0.772 (0.733, 0.811)	0.137 (0.134, 0.144)	
		Oversample	0.804 (0.768, 0.841)	0.168 (0.138, 0.170)	**<0.001***
		EyeDiff	0.817 (0.780, 0.846) ↑	0.168 (0.149, 0.178) ↑	**<0.001***
APTOS 2019 (*n* = 1100)	Severe Retinopathy (*n* = 58)	RETFound	0.867 (0.829, 0.906)	0.348 (0.323, 0.385)	
		Oversample	0.913 (0.884, 0.920)	0.372 (0.365, 0.409)	**<0.001***
		EyeDiff	0.914 (0.903, 0.934) ↑	0.496 (0.482, 0.504) ↑	**<0.001***
MESSIDOR2 (*n* = 526)	Proliferative Retinopathy (*n* = 11)	RETFound	0.960 (0.937, 0.988)	0.721 (0.712, 0.744)	
		Oversample	0.981 (0.952, 1.007)	0.724 (0.720, 0.741)	**<0.001***
		EyeDiff	0.980 (0.967, 0.994) ↑	0.740 (0.732, 0.769) ↑	**<0.001***
Glaucoma Fundus (*n* = 465)	Early Glaucoma (*n* = 87)	RETFound	0.860 (0.827, 0.873)	0.570 (0.561, 0.596)	
		Oversample	0.895 (0.866, 0.915)	0.791 (0.789, 0.807)	**<0.001***
		EyeDiff	0.927 (0.919, 0.934) ↑	0.809 (0.791, 0.846) ↑	**<0.001***
PAPILA (*n* = 98)	Glaucoma (*n* = 14)	RETFound	0.754 (0.742, 0.778)	0.391 (0.358, 0.424)	
		Oversample	0.772 (0.739, 0.773)	0.387 (0.347, 0.407)	**0.017***
		EyeDiff	0.795 (0.756, 0.813) ↑	0.401 (0.369, 0.407) ↑	**<0.001***

P values were compared between baseline RETFound and EyeDiff (EyeDiff), and between EyeDiff and Oversample (Oversample).
CI confidence interval, *AUROC* Area under the receiver operating characteristic curve, *AUPR* Area under the precision-recall curve.
*$P < 0.05$.

retinitis pigmentosa, choroidal osteoma, morning glory syndrome, incontinentia pigmenti-associated retinopathy, Von Hippel–Lindau syndrome, and cat scratch disease (see Table 6).

Discussion

In this study, we developed EyeDiff, a generative foundation model capable of synthesizing multimodal ophthalmic images from text prompts. Both quantitative evaluations and expert human assessments confirmed EyeDiff's ability to generate lesion-preserving images across multiple imaging modalities. When used as a data augmentation strategy, EyeDiff-generated images significantly enhanced diagnostic performance for underrepresented disease subtypes across 11 validation datasets. Importantly, these improvements were achieved on top of established modality-specific and multimodal foundation models, further boosting diagnostic accuracy for both common and rare ophthalmic conditions. These findings highlight EyeDiff's potential as a generalist model that enables the generation of diverse, clinically meaningful images from simple textual inputs. This approach addresses the persistent challenge of limited multimodal data, especially for rare conditions, and provides substantial value in enhancing expert-level diagnostic models through more balanced and comprehensive datasets.

Despite the rarity of individual rare diseases, their collective burden is considerable, impacting over 900 ocular abnormalities and often leading to lifelong vision impairment in many individuals[35]. While DL models hold promise for screening rare eye diseases, progress significantly lags behind that of common diseases due to the scarcity of large, annotated datasets[36,37]. Although ongoing efforts on foundation models like EyeFound and EyeCLIP have demonstrated potential in identifying rare diseases through learning from multimodal ophthalmic data[25,26], the class imbalance in real-world scenarios, due to the varying prevalence of different subtypes and the rarity of certain diseases, hinders training effectiveness and robustness. EyeDiff directly mitigates these challenges by generating lesion-preserving, modality-specific images from simple text prompts. Trained on aligned disease and manifestation labels across imaging modalities, EyeDiff learns how specific diseases present across different modalities and generates representative images that enhance the detection of rare diseases and minority disease subtypes. This functionality highlights its potential as a generalist model, particularly beneficial in contexts where data sharing is limited due to privacy concerns.

Generative AI (GenAI) has previously been applied in ophthalmology to generate images, but these efforts have largely been restricted to single-modality outputs[14,15,38–41]. Among GenAI methods, diffusion models have outperformed GANs in generating images with superior quality, diversity, stability, and controllability[42]. Diffusion models excel in generating high-fidelity images due to their iterative denoising process. While GANs may struggle with mode collapse (resulting in less diversity) and VAEs can suffer from blurry outputs, diffusion models are designed to progressively refine the generated images, leading to more realistic and high-resolution results. This makes them particularly suited for medical image generation, where fine details are crucial[20]. Unlike GANs, where the generator's output can sometimes be unpredictable, or VAEs that focus on a lower-dimensional latent space, diffusion models allow precise control over image characteristics by conditioning on various factors, such as labels or input attributes. This added control is a significant advantage in applications requiring customization, like generating specific medical conditions or anatomical structures.

EyeDiff achieves lesion preservation through a structured conditional generation mechanism. Specifically, the diffusion process is guided by multi-modal conditioning, where both textual and anatomical cues direct the network's attention to clinically relevant regions. Cross-attention ensures that lesion-related features are localized and faithfully reconstructed, rather than being freely hallucinated. During training, we adopted a data-driven training paradigm in which shared anatomical priors and disease-relevant lesion characteristics are learned implicitly through large-scale real-world ophthalmic data and text-guided diffusion modeling, thereby ensuring anatomical plausibility without hard-coded structural rules. This supervision, combined with fine-tuned control over the denoising trajectory, enforces consistency of pathological details across diverse cases. Although multiple strategies collectively provide practical safeguards against misleading artifacts, it remains challenging to fully eliminate hallucination and bias risk within a generative framework. Future work may incorporate automated artifact detection mechanisms or uncertainty-aware generation strategies to further quantify and control hallucination risk.

To further ensure reliability, EyeDiff's outputs are validated both quantitatively (e.g., downstream classification performance) and qualitatively (expert evaluation), confirming that generated lesions correspond to realistic and medically accurate features. Although some synthetic images

Table 6 | Comparison of the performance among the baseline RETFound model, oversample augmentation, and EyeDiff augmentation in rare disease classification tasks

Dataset	Disease	Model	AUROC	AUPR	*P*-value
Rare Diseases	Stargardt Disease	RETFound	0.695 (0.692, 0.717)	0.513 (0.489, 0.549)	
		Oversample	0.752 (0.714, 0.753)	0.42 (0.382, 0.44)	**<0.001***
		EyeDiff	0.816 (0.779, 0.828) ↑	0.723 (0.698, 0.745) ↑	**<0.001***
	Retinopathy of Prematurity	RETFound	0.733 (0.715, 0.757)	0.202 (0.176, 0.231)	
		Oversample	0.736 (0.71, 0.763)	0.247 (0.231, 0.249)	**0.045***
		EyeDiff	0.737 (0.734, 0.753) ↑	0.382 (0.352, 0.392) ↑	**<0.001***
	Retinoblastoma	RETFound	0.74 (0.712, 0.755)	0.316 (0.282, 0.33)	
		Oversample	0.712 (0.68, 0.724)	0.536 (0.519, 0.561)	**<0.001***
		EyeDiff	0.787 (0.761, 0.811) ↑	0.742 (0.719, 0.761) ↑	**<0.001***
	Retinitis Pigmentosa	RETFound	0.635 (0.612, 0.657)	0.442 (0.414, 0.445)	
		Oversample	0.637 (0.624, 0.672)	0.444 (0.43, 0.468)	**<0.001***
		EyeDiff	0.684 (0.672, 0.7) ↑	0.491 (0.489, 0.515) ↑	**<0.001***
	Choroidal Osteoma	RETFound	0.723 (0.703, 0.752)	0.216 (0.209, 0.23)	
		Oversample	0.589 (0.552, 0.596)	0.173 (0.164, 0.208)	**<0.001***
		EyeDiff	0.724 (0.722, 0.735)	0.256 (0.219, 0.295)	**0.056**
OphthalWeChat	Morning Glory Syndrome	RETFound	0.611 (0.584, 0.641)	0.198 (0.177, 0.215)	
		Oversample	0.639 (0.615, 0.667)	0.244 (0.218, 0.259)	**<0.001***
		EyeDiff	0.694 (0.659, 0.722) ↑	0.381 (0.356, 0.402) ↑	**<0.001***
	Incontinentia Pigmenti-Associated Retinopathy	RETFound	0.591 (0.565, 0.616)	0.071 (0.060, 0.082)	
		Oversample	0.627 (0.604, 0.653)	0.11 (0.098, 0.123)	**<0.001***
		EyeDiff	0.665 (0.641, 0.692) ↑	0.164 (0.149, 0.179) ↑	**<0.001***
	Von Hippel–Lindau syndrome	RETFound	0.648 (0.622, 0.672)	0.273 (0.256, 0.298)	
		Oversample	0.685 (0.663, 0.708)	0.319 (0.298, 0.337)	**<0.001***
		EyeDiff	0.731 (0.706, 0.753) ↑	0.444 (0.426, 0.465) ↑	**<0.001***
	Cat Scratch Disease	RETFound	0.628 (0.606, 0.654)	0.117 (0.105, 0.129)	
		Oversample	0.665 (0.641, 0.689)	0.16 (0.146, 0.173)	**<0.001***
		EyeDiff	0.706 (0.681, 0.729) ↑	0.234 (0.220, 0.248) ↑	**<0.001***
	Choroidal Osteoma	RETFound	0.643 (0.618, 0.666)	0.139 (0.126, 0.152)	
		Oversample	0.675 (0.652, 0.699)	0.185 (0.169, 0.200)	**<0.001***
		EyeDiff	0.722 (0.698, 0.748) ↑	0.268 (0.252, 0.284) ↑	**<0.001***

P values were compared between baseline RETFound and EyeDiff (EyeDiff), and between EyeDiff and Oversample (Oversample).
CI confidence interval, *AUROC* Area under the receiver operating characteristic curve, *AUPR* Area under the precision-recall curve.
*$P < 0.05$.

were recognizable due to minor artifacts, most retained key diagnostic features for downstream augmentations. Notably, all generated images were used in downstream model training without cherry-picking, demonstrating the robustness of lesion-related features and the model's practical utility as a reliable data source.

A key concern with generative models lies in their adaptability and whether generated images can effectively boost diagnostic performance. In our evaluation across eleven diverse datasets, EyeDiff augmented the underrepresented subtypes. Integration of these synthetic images led to notable performance improvements across different established foundation models, with statistically significant gains in AUROC and AUPR across datasets covering DR, glaucoma, vitreoretinal diseases, and rare conditions. EyeDiff demonstrated a consistent ability to mitigate biases arising from data imbalance, providing a scalable strategy for constructing more robust retinal diagnostic models through balanced data augmentation. While synthetic data offers opportunities to augment training datasets and improve medical AI systems, clear ethical guidelines on how synthetic data is validated and integrated into clinical workflows are warranted. Firstly, synthetic retinal images must be clearly labeled to distinguish them from real patient data, ensuring transparency in research and clinical use. Secondly, access to image-generation tools should be restricted to authorized personnel, with policies in place to prevent fraud or malicious use. Finally, developers should monitor and audit the use of these tools regularly to mitigate misuse, such as creating misleading medical content.

Nonetheless, this study has several limitations. First, the current training data still lacks sufficient population diversity, which may affect the generalizability of the model. Future work incorporating prevalence-stratified analyses across broader epidemiological distributions may further clarify the generalizability of synthetic augmentation strategies. Second, some generated images still exhibit noticeable visual differences from real data. These discrepancies may introduce risks of diagnostic bias, such as deviations in lesion location or color. To address these challenges, expanding the dataset to include more diverse, real-time data and fostering closer collaboration between clinical and engineering teams will be crucial for enhancing image quality and validating the clinical utility of synthetic images. Moreover, the current text prompts used were relatively simplified; future work should focus on developing algorithms capable of handling more complex, nuanced prompts while ensuring fairness and eliminating bias in prompts.

In conclusion, EyeDiff is a generative foundation model capable of synthesizing multimodal ophthalmic images from natural language prompts. Integrating EyeDiff-generated images with real-world data

significantly enhanced the diagnostic accuracy of the baseline foundation model across both common and rare ophthalmic diseases. By enabling efficient, high-quality, and balanced data augmentation, EyeDiff provides a practical solution to the long-standing challenge of data scarcity, particularly for underrepresented disease subtypes. This work lays a foundation for building generalizable and robust disease detection models. Future efforts should focus on curating large-scale, diverse image–text datasets from multiple regions and refining generation algorithms to further improve image fidelity, semantic alignment, and controllability.

Methods

Data acquisition

EyeDiff was developed using a large-scale collection of ophthalmic image-text pairs, covering 14 modalities and over 80 disease categories from 8 datasets. This extensive training enabled the model to accurately learn the associations between image distributions and their corresponding textual descriptions across a wide range of diseases. We then assessed the generalizability of EyeDiff on 2 internal datasets and 9 external datasets for several common and rare disease classifications, as well as minority class augmentation. Table 1 summarizes the characteristics of the training and validation datasets, and the study flowchart is illustrated in Fig. 1. This study utilized publicly available data and was approved by the Institutional Review Board of the Hong Kong Polytechnic University (HSEARS20240202004).

Training datasets

A brief overview of the eight training datasets is provided below: (1) Multimodal datasets: The Retinal Image Bank and GAMMA were included. The Retinal Image Bank, established by the American Society of Retinal Specialists, is an open-access collection containing over 29,000 multimodal images and descriptions covering various retinal diseases (Supplementary Table 3). GAMMA is a multimodal image dataset designed for glaucoma grading. It consists of fundus and OCT images from 300 patients, annotated with glaucoma grade, macular fovea coordinates, and optic disc/cup segmentation masks[43]. (2) CFP datasets: The following datasets were used: EyePACS, REFUGE, ORIGA, RIM-ONE, and DRISHTI. EyePACS comprises 88,702 fundus images from a diverse population with varying degrees of DR and is widely used to develop and evaluate DR screening models[44]. REFUGE provides 1,200 fundus images with ground truth segmentations and glaucoma labels, making it the largest publicly available dataset for automated glaucoma assessment[45]. ORIGA contains 650 fundus images annotated by clinical experts, with a focus on disc and cup segmentation and the cup-to-disc ratio (CDR) estimation[46]. RIM-ONE offers 485 fundus images (313 from healthy individuals and 172 from glaucoma patients) for optic nerve evaluation, each manually segmented by a glaucoma specialist[47]. DRISHTI consists of 101 fundus images from both normal and glaucomatous eyes, with optic disc and cup segmented by four experts[48]. (3) OCT datasets: The OCTDL dataset was used, consisting of over 2,000 OCT images labeled with disease groups and retinal pathologies. These include AMD, diabetic macular edema (DME), epiretinal membrane (ERM), retinal artery occlusion (RAO), RVO, and common vitreomacular interface diseases[32].

Validation datasets

The generalizability of EyeDiff was validated using 11 open-access ophthalmic image datasets, with 2 datasets for internal validation (OCTDL and Retina Image Bank) and 9 datasets for external validation (APTOS 2019, OphthalWeChat, MESSIDOR-2, Glaucoma Fundus, JSIEC, Retina, OCTID, IDRiD, and PAPILA).

DR diagnosis: The APTOS-2019 (India), IDRiD (India), and MESSIDOR-2 (France) datasets were used. These datasets contain color fundus images labelled according to the International Clinical Diabetic Retinopathy Severity Scale, which classifies disease into five stages: no DR, mild non-proliferative DR (NPDR), moderate NPDR, severe NPDR, and proliferative DR.

Glaucoma diagnosis: The PAPILA (Spain) and Glaucoma Fundus (South Korea) datasets were used, both of which provide color fundus images annotated with three categories: non-glaucoma, early glaucoma (suspected glaucoma), and advanced glaucoma.

Multi-category eye disease diagnosis: The JSIEC (China), Retina, OCTID, OCTDL, and OphthalWechat datasets were applied. JSIEC comprises 1000 color fundus images, covering 39 common fundus diseases and conditions. The Retina dataset contains 601 color fundus images, categorized as normal, glaucoma, cataract, and retinal diseases. OCTID includes 572 OCT scans labeled as normal, macular hole, AMD, central serous chorioretinopathy (CSCR), and DR. OCTDL includes OCT images annotated with conditions such as normal, AMD, DME, ERM, RVO, etc. OphthalWeChat consists of 3071 ophthalmic images and corresponding captions that were published on an open-access WeChat Official Account, which represents a publicly accessible platform that allows verified users to share text and image-based medical content. The images used in the current study were collected from January 1, 2016, to December 31, 2024.

Rare disease diagnosis: Multimodal images collected from rare diseases in the Retinal Image Bank between 2019 and 2023, as well as the OphthalWeChat dataset, were used. We created a custom dictionary to identify rare disease cases, such as birdshot chorioretinopathy, cone dystrophy, Stargardt disease, etc. Rare disease categories can be retrieved from authoritative databases, including the American Academy of Ophthalmology, Orphanet, and the National Organization for Rare Disorders. A total of 5749 images representing 17 rare diseases were included. The details of each dataset, including the imaging modality, examination device, disease sub-categories, and the distribution of images for training, validation, and testing, are presented in Supplementary Table 4.

Well-designed text prompt construction

The textual prompts were constructed through an information fusion process across multiple training datasets. Specifically, we extracted disease-related annotations and corresponding imaging modality from the original dataset labels and standardized them using a custom-built dictionary. The dictionary itself was developed with clinical input to ensure medical accuracy and consistency. The final prompts consisted of structured components, including imaging modality, disease or lesion type, and disease severity (such as mild NPDR, moderate NPDR, etc.) when available. A detailed list of prompt examples is provided in Supplementary Table 5. We excluded non-routine retinal examination image descriptions, such as histology and pathology images.

Algorithm architecture

Diffusion models are probabilistic generative models that learn data distributions by progressively denoising a variable initially sampled from a normal distribution. The latent diffusion model (LDM) extends this concept by incorporating a compression approach. Specifically, it compresses images into a more efficient, lower-dimensional latent space, which is then used for the diffusion process. Features of textual information interact with the features during the image generation process, ultimately enabling conditional generation. Compared to the high-dimensional pixel space, the compact latent space offers several advantages for likelihood-based generative models. First, it enables the model to concentrate on the most meaningful semantic features of the data. Second, it allows for training in a much lower-dimensional space, which significantly reduces computational complexity, making the model both more efficient and scalable.

The architecture of EyeDiff is built upon the Stable Diffusion v1.5 (SD), a text-to-image generation model implemented based on LDM. Among existing text-to-image models, SD has garnered significant attention for its impressive performance in generating high-quality images and its cost-effective fine-tuning. Its denoising process operates in a latent space, akin to other diffusion models, producing images that are highly consistent with the input text. This makes SD particularly suitable for text-guided image generation[20,49]. Fig. 1b presents the algorithm architecture of EyeDiff. It is formed by two modules: one is a Variational Autoencoder (VAE), which

learns a compressed latent space, and the second is an LDM, which forms the core component for text-to-image generation. The encoder-decoder framework is pre-trained separately to establish an effective latent representation of the input images before training the diffusion model. Importantly, the encoder and decoder are usually kept frozen during diffusion model training to maintain a stable and consistent latent space.

Model development

For training EyeDiff, we utilized structured text prompts as input and paired images as the ground truth. Firstly, the VAE model is trained to project the image data (resized to 512 × 512 resolution) into two latent tensors: a mean and a standard deviation. These latent tensors are used to define a Gaussian distribution, which is sampled to produce compact embeddings of the image data. From this compact embedding, the VAE decoder reconstructs the input image. The VAE is trained by using a combination of L1 reconstruction loss, learned perceptual image patch Similarity (LPIPS)[50], and a Kullback-Leibler (KL)-divergence loss. The KL-divergence loss ensures a Gaussian distribution in the latent space, while the integration of the remaining loss functions avoids the blurring effects typically associated with relying exclusively on pixel-space losses like L1 objectives. Once the VAE is trained, we use it to turn images into latent representations that capture essential visual features. The LDM then works on this latent space.

During the training phase of the LDM, noise is gradually added to the image latent according to the linear noise scheduler at each timestep. The U-Net of the latent diffusion model then learns to reduce this noise by taking as input the noisy image latent, the timestep information, and the text prompts. The predicted noise is subtracted from the noisy image latent at each timestep, resulting in a denoised latent representation at the end of the process, which can be passed through the VAE decoder to obtain the generated image. For the conditional mechanism, the text prompts are encoded through a contrastive language-image pre-training (CLIP) text encoder and fed into the denoising U-Net of the LDM via cross-attention. The training loss function serves as a reconstruction objective, comparing the noise added to the latent representation with the predictions generated by the UNet. For generating a new image, the process is the same, but we start from total random noise until we have a synthetic image whose generation has been guided by the text embedding.

For implementation details, we utilized the AdamW optimizer with a weight decay of 1e-2. The batch size was set to 8, and learning rate to 5e-5. Each training session was preset to run for a total of 5 epochs, given that each epoch involved training with sufficiently comprehensive data.

Objective evaluation of text-to-image alignment

We applied VQAScore[27] to objectively measure the alignment between generated images and the corresponding texts in downstream tasks. The VQAScore applied a visual-question-answering (VQA) model to compute an alignment score by estimating the probability of a "Yes" answer to the question "Does this figure show '[text]'?". The VQAScore offers a simple yet effective approach and has been shown to outperform prior art[51], demonstrating strong agreement with human judgements. This metric ranges from 0 to 1, with higher scores indicating better alignment.

Subjective evaluation of medical realism

We applied the Turing test and visual quality scores to evaluate the medical realism of the generated images. Two ophthalmologists (board-certified ophthalmologists R.C. and X.C., both with 5 years of experience) independently participated in Turing test and visual quality evaluation.

Turing test: One hundred images were randomly selected for the Turing test, with 50 real images and 50 generated images. Two ophthalmologists were asked to independently determine whether each image was a real clinical image or a synthetic one. The accuracy of ophthalmologists in correctly identifying real and generated images was calculated to evaluate the perceptual realism of the generated outputs[52].

Visual quality evaluation: Fifty generated images were randomly selected for visual quality evaluation. The ophthalmologists evaluated the generated images subjectively using a five-point scale, based on the integrity of generated structures and lesions according to text prompts. The scale is as follows: 1 = The modality and lesion features of the generated image fully match the text prompts; 2 = The modality of the generated image corresponds to the text prompts, and the lesion features mostly align with the text prompts; 3 = The modality of the generated image corresponds to the text prompts, and the lesion features slightly align with the text prompts; 4 = The modality of the generated image corresponds to the text prompts, but the lesion features cannot be generated; 5 = All the text-guided features cannot be generated. To determine the inter-rater agreement, we calculated Cohen's weighted kappa score. This score ranges from −1 to 1, with values between 0.40 and 0.60 indicating moderate agreement, 0.60 and 0.80 substantial agreement, and 0.80 to 1.00 representing almost perfect agreement.

Downstream evaluation of applicability

We evaluated the applicability of EyeDiff-generated images in augmenting minority classes and enhancing the overall performance of disease diagnosis on the validation datasets listed above. We used RETFound, EyeCLIP, and EyeFound, a series of established foundation model for retinal disease diagnosis, as the baseline models for our experiments[24]. The weight sets of RETFound, EyeFound, and EyeCLIP were applied, respectively.

RETFound and EyeFound were developed based on a Vision Transformer (ViT) architecture. It was utilized to extract image features into a 1024-dimensional embedding, followed by a multi-head attention mechanism and a fully connected layer to obtain the diagnosis output. The ViT is capable of capturing long-range dependencies and global contextual information by leveraging a self-attention mechanism to process image patches[53]. Its ability to effectively handle non-local information makes it a powerful model for medical image analysis, where diagnoses often depend on global features[54]. Previous evidence has confirmed the effectiveness of ViT in detecting lesions and classifying retinal diseases from multimodal ophthalmic images[55,56].

EyeCLIP was developed based on the CLIP framework, which learns a shared embedding space for image–text pairs through contrastive learning. To enhance visual representation learning, an image decoder inspired by the Masked Autoencoder (MAE) structure is integrated into the CLIP architecture, enabling masked image reconstruction in addition to the original contrastive objectives. The training objective combines image-text contrastive loss, image-image contrastive loss for cross-modality alignment, and reconstruction loss for self-supervised feature enhancement. All image modalities share a unified vision encoder, allowing the model to learn modality-invariant representations without the need for explicit fusion layers. During training, images are cropped to the field of view, resized to 224 × 224, and augmented with random cropping, color jitter, and horizontal flipping.

We compared the diagnostic performance in the following experiments: (1) RETFound/EyeCLIP/EyeFound: the baseline diagnostic model using original real images; (2) Oversample: the diagnostic model using original real images and oversampled images for augmenting minority classes; (3) EyeDiff: the diagnostic model using original real images and EyeDiff-generated images for augmenting minority classes. All classes with fewer samples than the majority class in the dataset were augmented using oversampling (in the oversample group) or EyeDiff (in the EyeDiff group) to ensure balance by matching the number of samples in the majority class. For single-modality diagnostic tasks (e.g. color fundus images or OCT images), EyeDiff was prompted with text such as "fundus image, [disease category]" or "optical coherence tomography, [disease category]" to generate modality-specific images with corresponding lesions. These images were then used to augment the minority classes in developing classification models. For multi-modality diagnostic tasks, the data used to train EyeDiff was labelled with disease and manifestation information aligned to each specific modality. This approach enables the model to learn the distinct manifestations of each disease across modalities and generate representative images for downstream tasks.

The performance of retinal disease diagnosis was evaluated through standard metrics, including the AUROC and the AUPR.

Data availability

The data for model training in the current study are available as open data through the following links: Retinal Image Bank (https://imagebank.asrs.org/), EyePACS (https://www.kaggle.com/c/diabetic-retinopathy-detection/data), OCTDL (https://ieee-dataport.org/documents/octdl-optical-coherence-tomography-dataset-image-based-deep-learning-methods), REFUGE (https://bitbucket.org/woalsdnd/refuge/src/master/), ORIGA (https://figshare.com/articles/dataset/Retinal_Fundus_Glaucoma_Image_dataset/24549217?file=43119880), RIM-ONE (https://bit.ly/rim-one-dl-images), DRISHTI (https://www.kaggle.com/datasets/lokeshsaipureddi/drishtigs-retina-dataset-for-onh-segmentation), GAMMA (https://paperswithcode.com/dataset/gamma-challenge). The data for validation in the current study are available as open data through the following links: IDRID (https://ieee-dataport.org/open-access/ indian-diabetic-retinopathy-image-dataset-idrid), MESSIDOR-2 (https://www.adcis.net/en/third-party/messidor2/), APTOS-2019 (https://www.kaggle.com/competitions/aptos2019-blindness-detection/data), PAPILA (https://figshare.com/articles/dataset/PAPILA/14798004/1), Glaucoma Fundus (https://dataverse.harvard.edu/dataset. xhtml?persistentId=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) and OCTDL(https://ieee-dataport.org/documents/octdl-optical-coherence-tomography-dataset-image-based-deep-learning-methods).

Code availability

The deep-learning model was developed using PyTorch (http://pytorch.org). We trained the model on an NVIDIA V100 card. The code for deep learning model development can be accessed at https://github.com/huggingface/diffusers/tree/main/examples/dreambooth.

Received: 28 October 2024; Accepted: 8 March 2026;
Published online: 24 March 2026

References

1. Raimundo, R. & Rosário, A. The impact of artificial intelligence on data system security: a literature review. *Sensors* **21**, https://doi.org/10.3390/s21217029 (2021).
2. Lama, H. et al. Severe macular complications in glaucoma: high-resolution multimodal imaging characteristics and review of the literature. *BMC Ophthalmol.* **23**, 318 (2023).
3. Stino, H. et al. Association of diabetic lesions and retinal nonperfusion using widefield multimodal imaging. *Ophthalmol. Retin.* **7**, 1042–1050 (2023).
4. Rahman, N., Georgiou, M., Khan, K. N. & Michaelides, M. Macular dystrophies: clinical and imaging features, molecular genetics and therapeutic options. *Br. J. Ophthalmol.* **104**, 451–460 (2020).
5. Vij, R. & Arora, S. A systematic review on diabetic retinopathy detection using deep learning techniques. *Arch. Comput. Methods Eng.* **30**, 2211–2256 (2023).
6. Vij, R. & Arora, S. A systematic survey of advances in retinal imaging modalities for Alzheimer's disease diagnosis. *Metab. Brain Dis.* **37**, 2213–2243 (2022).
7. Aung, Y. Y. M., Wong, D. C. S. & Ting, D. S. W. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. *Br. Med. Bull.* **139**, 4–15 (2021).
8. Gichoya, J. W. et al. AI recognition of patient race in medical imaging: a modelling study. *Lancet Digital Health* **4**, e406–e414 (2022).
9. Vij, R. & Arora, S. A novel deep transfer learning based computerized diagnostic Systems for Multi-class imbalanced diabetic retinopathy severity classification. *Multimed. Tools Appl.* **82**, 34847–34884 (2023).
10. Khalifa, N. E., Loey, M. & Mirjalili, S. A comprehensive survey of recent trends in deep learning for digital images augmentation. *Artif. Intell. Rev.* **55**, 2351–2377 (2022).
11. Goceri, E. Medical image data augmentation: techniques, comparisons and interpretations. *Artif. Intell. Rev*. 1-45 (2023).
12. Gao, L., Zhang, L., Liu, C. & Wu, S. Handling imbalanced medical image data: a deep-learning-based one-class classification approach. *Artif. Intell. Med.* **108**, 101935 (2020).
13. Khan, A. A., Chaudhari, O. & Chandra, R. A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and evaluation. *Expert Syst. Appl.* **244**, 122778 (2024).
14. Chen, R. et al. Translating color fundus photography to indocyanine green angiography using deep-learning for age-related macular degeneration screening. *NPJ Digit. Med.* **7**, 34 (2024).
15. 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).
16. Kugelman, J. et al. Data augmentation for patch-based OCT chorio-retinal segmentation using generative adversarial networks. *Neural Comput. Appl*. **33**, 7393–7408 (2021).
17. Yoo, T. K., Choi, J. Y. & Kim, H. K. Feasibility study to improve deep learning in OCT diagnosis of rare retinal diseases with few-shot classification. *Med. Biol. Eng. Comput.* **59**, 401–415 (2021).
18. Sonmez, S. C., Sevgi, M., Antaki, F., Huemer, J. & Keane, P. A. Generative artificial intelligence in ophthalmology: current innovations, future applications and challenges. *Br. J. Ophthalmol.* **108**, 1335–1340 (2024).
19. Chen, R. et al. Noninvasive synthesis of multiframe ultra-widefield fluorescein angiography from color fundus photographs. *Ophthalmol. Retina* https://doi.org/10.1016/j.oret.2025.08.002 (2025).
20. Rombach, R. et al. High-resolution image synthesis with latent diffusion models. 10674–10685 (2021).
21. Tian, Y., Fan, L., Isola, P., Chang, H. & Krishnan, D. J. A. StableRep: synthetic images from text-to-image models make strong visual representation learners. abs/2306.00984 (2023).
22. Xu, K. et al. Digital twins in ophthalmology: Concepts, applications, and challenges. *Asia Pac. J. Ophthalmol.* **14**, 100205 (2025).
23. Wu, X. et al. Generation of Fundus fluorescein angiography videos for health care data sharing. *JAMA Ophthalmol.* https://doi.org/10.1001/jamaophthalmol.2025.1419 (2025).
24. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
25. Shi, D. et al. EyeFound: a multimodal generalist foundation model for ophthalmic imaging. *ArXiv* **abs/2405.11338** (2024).
26. Shi, D. et al. A multimodal visual-language foundation model for computational ophthalmology. *NPJ Digit. Med.* **8**, 381 (2025).
27. Lin, Z. et al. in *Computer Vision–ECCV 2024*. (eds Aleš Leonardis *et al*.) 366-384 (Springer Nature Switzerland).
28. Porwal, P. et al. IDRiD: diabetic retinopathy–segmentation and grading challenge. *Med. Image Anal.* **59**, 101561 (2020).
29. 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).
30. 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).
31. Gholami, P., Roy, P., Parthasarathy, M. K. & Lakshminarayanan, V. OCTID: optical coherence tomography image database. *Comput. Electr. Eng.* **81**, 106532 (2020).
32. Kulyabin, M. et al. OCTDL: optical coherence tomography dataset for image-based deep learning methods. *Sci. Data* **11**, 365 (2024).
33. 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).
34. Xu, P. et al. Benchmarking large multimodal models for ophthalmic visual question answering with OphthalWeChat. *Adv. Ophthalmol. Pract. Res.* **6**, 33–41 (2025).

35. Sharma, M. Overcoming challenges in research and development of rare eye diseases. *Indian J. Ophthalmol.* **70**, 2214–2215 (2022).
36. Vij, R. & Arora, S. A Systematic Review on Deep Learning Techniques for Diabetic Retinopathy Segmentation and Detection Using Ocular Imaging Modalities. *Wirel. Personal. Commun.* **134**, 1153–1229 (2024).
37. Vij, R. & Arora, S. A hybrid evolutionary weighted ensemble of deep transfer learning models for retinal vessel segmentation and diabetic retinopathy detection. *Comput. Electr. Eng.* **115**, 109107 (2024).
38. He, S. et al. Bridging the camera domain gap with image-to-image translation improves glaucoma diagnosis. *Transl. Vis. Sci. Technol.* **12**, 20–20 (2023).
39. 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).
40. 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).
41. Zhang, W. et al. in *Medical Image Computing and Computer Assisted Intervention–MICCAI*. 689-699 (Springer Nature Switzerland).
42. Dhariwal, P. & Nichol, A. J. A. Diffusion models beat GANs on image synthesis. (2021).
43. Wu, J. et al. GAMMA Challenge: glaucoma grading from multi-modality images. **90**, 102938 (Elsevier, 2022).
44. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*. **316**, 2402–2410 (2016).
45. 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).
46. Zhang, Z. et al. ORIGA(-light): an online retinal fundus image database for glaucoma analysis and research. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. *IEEE Engineering in Medicine and Biology Society. Annual International Conference* **2010**, 3065-3068 (2010).
47. Fumero, F., Alayón, S., Sánchez, J. L., Sigut, J. F. & Gonzalez-Hernandez, M. J. t. I. S. o. C.-B. M. S. RIM-ONE: an open retinal image database for optic nerve evaluation. 1-6 (2011).
48. Sivaswamy, J. et al. Drishti-GS: retinal image dataset for optic nerve head(ONH) segmentation. 53-56 (2014).
49. Ho, J. Classifier-Free Diffusion Guidance. *ArXiv* (2022).
50. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 586-595 (IEEE, 2018).
51. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R. & Choi, Y. CLIPScore: a reference-free evaluation metric for image captioning. *ArXiv* **abs/2104.08718** (2021).
52. Moon, H. H. et al. Generative AI in glioma: ensuring diversity in training image phenotypes to improve diagnostic performance for IDH mutation prediction. *Neuro Oncol.* **26**, 1124–1135 (2024).
53. Al-Hammuri, K., Gebali, F., Kanan, A. & Chelvan, I. T. Vision transformer architecture and applications in digital health: a tutorial and survey. *Vis. Comput. Ind. Biomed. Art.* **6**, 14 (2023).
54. Aburass, S., Dorgham, O., Al Shaqsi, J., Abu Rumman, M. & Al-Kadi, O. Vision transformers in medical imaging: a comprehensive review of advancements and applications across multiple diseases. *J. Imaging Inform. Med.* https://doi.org/10.1007/s10278-025-01481-y (2025).
55. Rodriguez, M. A., AlMarzouqi, H. & Liatsis, P. Multi-label retinal disease classification using transformers. *IEEE J. Biomed. Health Inf.* **27**, 2739–2750 (2023).
56. Oulhadj, M. et al. Diabetic retinopathy prediction based on vision transformer and modified capsule network. *Comput Biol. Med.* **175**, 108523 (2024).

Acknowledgements

We thank the American Society of Retina Specialists for providing the valuable Retina Image Bank and the InnoHK HKSAR Government for providing valuable support. The study was supported by the Start-up Fund for RAPs under the Strategic Hiring Scheme (P0048623) from HKSAR, Global STEM Professorship Scheme (P0046113), and Henry G. Leong Endowed Professorship in Elderly Vision Health. The sponsors or funding organizations had no role in the design or conduct of this research.

Author contributions

D.S. conceived the study. D.S. built the deep learning model. D.S., R.C, and W.Z. conducted the literature search and analyzed the data. R.C. and X.C. completed human evaluation. W.Z. performed validation of downstream tasks and quantitative evaluation. R.C. wrote the manuscript. R.C, B.L, P.X., S.L, and X.W. organized figures and tables in this study. M.H. provided the data and facilities. All authors 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/s41746-026-02560-2.

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

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) 2026