This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

RetinaLogos: Fine-Grained Synthesis of High-Resolution Retinal Images Through Captions

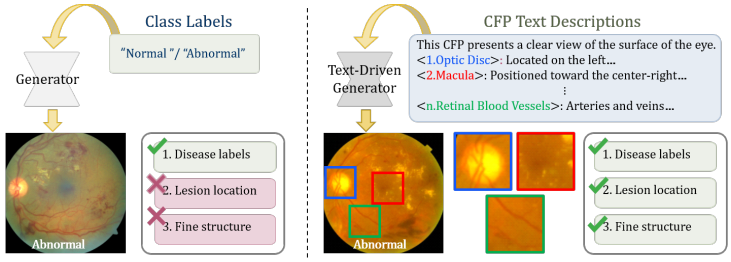
Junzhi Ning[*1], Cheng Tang[*1,3], Kaijing Zhou[4], Diping Song[1], Lihao Liu[1], Ming Hu[1,5], Wei Li[6], Huihui Xu[1], Yanzhou Su[7], Tianbin Li[1], Jiyao Liu[8], Jin Ye[5,1], Sheng Zhang[9], Yuanfeng Ji[10], Junjun He[1,2†]

[1] Shanghai Artificial Intelligence Laboratory, China
[2] Shanghai Innovation Institute, China
[3] Shanghai Institute of Laser Technology, China
[4] Eye Hospital, Wenzhou Medical University, China
[5] Monash University, Australia
[6] Shanghai Jiao Tong University, China
[7] Fuzhou University, China
[8] Fudan University, China
[9] Imperial College London, United Kingdom
[10] Stanford University, USA
hejunjun@pjlab.org.cn

Abstract. The scarcity of high-quality, labelled retinal imaging data, which presents a significant challenge in the development of machine learning models for ophthalmology, hinders progress in the field. Existing methods for synthesising Colour Fundus Photographs (CFPs) largely rely on predefined disease labels, which restricts their ability to generate images that reflect fine-grained anatomical variations, subtle disease stages, and diverse pathological features beyond coarse class categories. To overcome these challenges, we first introduce an innovative pipeline that creates a large-scale, captioned retinal dataset comprising 1.4 million entries, called *RetinaLogos-1400k*. Specifically, *RetinaLogos-1400k* uses the visual language model (VLM) to describe retinal conditions and key structures, such as optic disc configuration, vascular distribution, nerve fibre layers, and pathological features. Building on this dataset, we employ a novel three-step training framework, called *RetinaLogos*, which enables fine-grained semantic control over retinal images and accurately captures different stages of disease progression, subtle anatomical variations, and specific lesion types. Through extensive experiments, our method demonstrates superior performance across multiple datasets, with 62.07% of text-driven synthetic CFPs indistinguishable from real ones by ophthalmologists. Moreover, the synthetic data improves accuracy by 5%-10% in diabetic retinopathy grading and glaucoma detection. Codes are available at Link.

Keywords: Retinal Imaging · Text-to-Image Generation · Medical Data Synthesis · Fine-grained Controllable generation

[*] Equal contribution.
[†] Corresponding author.



(a) CFP Generation with Disease Labels (Previous) (b) CFP Generation Guided via Fine-Grained Captions (Ours)

Fig. 1. Class-Conditioned CFP Generation vs. Text-Driven CFP Generation.

1 Introduction

Eye healthcare has become a major global concern, as untreated ocular conditions can severely impact an individual's quality of life [4]. Many people have difficulty accessing ophthalmic resources, particularly in resource-limited areas [25]. To overcome the challenges posed by limited ophthalmic resources, early detection of eye diseases is crucial, as it enables timely intervention and can help prevent irreversible vision loss [14]. Among the available diagnostic tools, non-invasive fundus imaging, particularly Color Fundus Photography (CFP), is one of the most widely used and affordable methods in daily clinical practice. Recent advancements in deep learning have significantly transformed the field, enabling the automated analysis of CFP and offering great promise for the early detection of common eye diseases [22,15]. Current deep-learning techniques rely heavily on large-scale datasets to train various downstream models for CFP. For instance, training foundational CFP models [31,5,24] that are competent in zero-shot downstream tasks requires at least a million-level dataset to facilitate model convergence. Despite these advancements, the scarcity of CFP, both in quantity and quality, emphasizes the urgent need for more high-quality data in this domain.

Generative models that synthesize data for training various downstream medical tasks have shown significant success [2,18,11,28,10,17], and provide a feasible solution to address the issues of data scarcity. For instance, methods in [29,30,20,19,13] used Generative Adversarial Networks conditioned on features such as blood vessel structure, lesion region masks, and disease labels to generate retinal images. A two-stage approach has also been adopted in [8,1], in which the first stage generates realistic conditions, and the second stage generates retinal images based on these conditions. However, current generative methods [23,8] primarily rely on the conditions of predefined disease labels, which restrict the generated images to broader categories with diverse anatomical structures. As shown in Fig. 1, this limitation prevents the generation of CFP with more fine-grained details—such as varying stages of retinal disease, subtle anatomical variations, or specific lesion types.

To address the above challenges, we first introduce a data collection pipeline designed to amass a large-scale captioned caption dataset totalling 1.4 million real CFPs paired with synthetic detailed captions, which are sourced from both open-source and private datasets. Leveraging this extensive dataset, we then propose *RetinaLogos*, a novel text-to-image framework for retinal image synthesis. Specifically, using these extensive text-retinal image pairs, we then develop a tailored text-to-image generator capable of not only synthesizing high-resolution retinal images but also offering fine-grained control over specific anatomical structures and disease progressions. Our method allows for the generation of diverse, visually plausible synthetic CFP, where the appearance can be manipulated through free-form descriptions and prompts.

In summary, our main contributions are as follows: a) We propose a comprehensive data collection pipeline, which assembles what is currently the largest **1.4 million** captioned CFP dataset (1.4 million CFPs paired with synthetic captions) to support advancements in text-driven retinal image synthesis. b) We propose *RetinaLogos*, a novel text-to-image framework for retinal image generation. *To the best of our knowledge, RetinaLogosis the first to explore large-scale generation of CFPs from textual descriptions, supported by a dataset exceeding one million CFPs-caption pairs.* c) Our method achieves state-of-the-art performance in text-driven CFP synthesis, demonstrating superior fidelity and clinical relevance on the EyePACS, REFUGE, and IROGS datasets. This has been validated through improved Frechet Inception Distance (FID) and Retina CLIP scores, as well as through evaluations based on criteria defined by expert ophthalmologists.

2 Proposed Methodology

2.1 Retinal Captioning and Data Synthesis Pipeline

Authentic Image Quantity & **Diversity.** The scale and diversity of the dataset are critical factors influencing the performance of generative models. We constructed a comprehensive dataset of CFPs and corresponding captions, comprising over 1.4 million real-world fundus images sourced from both open-access and private datasets. This dataset includes both images and their corresponding Electronic Health Records (EHRs), as illustrated in Fig. 2(a). The fundus images span a broad spectrum of retinal diseases, while the associated EHRs provide essential information to provide grounded labels, including subclinical disease labels, disease severity ratings, and the general health status of patients. Additionally, the EHRs contain diagnostic reports contributed by healthcare professionals.

Caption Generation & **Reliability.** As shown in Fig. 2(a), captions are generated using a powerful VLM with the CFP and its corresponding EHR as multimodal inputs. To be more specific, in our proposed data construction pipeline, the VLM is prompted to function as a professional retinal imaging expert, denoted as model $\mathcal{E}$, to generate detailed descriptions based on diagnostic

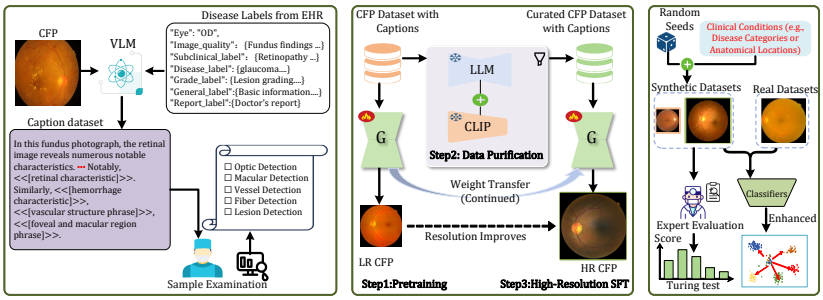


Fig. 2. *RetinaLogos* architecture overview. (a) EHR text and CFP images are integrated via a vision-language model for clinical caption generation. (b) Multi-resolution CFP synthesis includes low-resolution generation, LLM-guided data purification, and high-resolution fine-tuning. (c) Generated CFPs are tested for authenticity, disease classification, and expert evaluation.

symptomatology extracted from EHR and the corresponding paired fundus images. Let $\{T_i\}_{i=1}^N$ represent the EHR data and $\{D_i^l\}_{i=1}^N$ denote the corresponding paired fundus images. The captions generation process can be formally formulated as follows:

$$C_i^l = \mathcal{E}(T_i, D_i^l), \quad i = 1, 2, \ldots, N. \tag{1}$$

Furthermore, to ensure that the captions align with clinical expectations, professional ophthalmologists are involved in reviewing the generated annotations, represented as paired text-to-image data $\{D_i^l, C_i^l\}_{i=1}^N$.

2.2 Retinal Image Synthesis via Text-to-Image Generation Framework

In this study, we trained a latent flow-matching DIT model inspired by [7], as our retinal text-to-image generator. We employed the frozen Google Gemma 2B [26] text decoder to obtain the word embeddings of the retinal image captions and leveraged the flow matching mechanism to linearly interpolate between noise and the clean sample. Mathematically, given an CFP sample $x^* \sim p_{data}$, an associated caption ϕ, and $\epsilon \sim \mathcal{N}(0, I)$, the linear interpolation forward process is formulated as $x_t = \alpha_t x^* + \beta_t \epsilon = t x^* + (1-t)\epsilon$, where $t \in [0, 1]$. Its corresponding vector field is $v_t(x_t) = x^* - \epsilon$. During training, the model is optimized using the following conditional flow-matching objective:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t\sim U(0,1), x^*\sim p_{data}, \epsilon\sim\mathcal{N}(0,\mathbf{I})} \left[\left\| v_\theta\left(x_t, t, \phi\right) - v_t(x_t) \right\|_2^2 \right] \tag{2}$$

The training process consisted of three steps, as outlined in Fig.2(b).

Step I: Pretraining Stage. The pretraining stage initialised the model weights using the checkpoint from [32] as a starting point, the warm-up approach reduced

training time by closing the gap between the visual representations of natural and retinal images. The primary objective of this stage was to effectively adapt the model's backbone for the retinal generation task. The model was pre-trained on 1.4 million retinal images, where the corresponding annotated descriptions are denoted as $\{C_i^l\}_{i=1}^N$, at a low resolution of 256×256, with the corresponding images represented as $\{D_i^l\}_{i=1}^N$.

Step II: Precision Filtering and Semantic Refinement. Ensuring data quality is indispensable for training text-driven image generation models for CFPs, we implement two strategies to further purify the datasets of image-text pairs. Firstly, we employ the existing RetinClip encoders [24] to sift through retinal images and caption annotations, filtering out pairs with CLIP similarity scores below 0.6, an operation captured by:

$$\{D_i^h, C_i^h\}_{i=1}^N = \{(D_i^l, R(C_i^l)) \mid S(D_i^l, C_i^l) \geq 0.6 \; s.t. \; i \in \{1, 2, 3..., N\}\}, \quad (3)$$

where $S(D_i^l, C_i^l)$ quantifies the semantic alignment between an image D_i^l and a caption C_i^l, and $R(C_i^l)$ denotes the refined caption. The dataset before filtering and refinement is denoted as $\{D_i^h, C_i^h\}_{i=1}^N$, which includes all collected pairs retinal images D_i^l and captions C_i^l. After filtering, the resulting dataset $\{D_i^h, C_i^h\}_{i=1}^N$ contains only the image-caption pairs with sufficient semantic alignment. Then, we further refined the captions with the Qwen 2.5 LLM [27] using designed medical prompts. The prompt is carefully tailored to eliminate unnecessary descriptions prevalent in the retinal image content (such as recommendations to avoid liability) while preserving the original meaning of the captions.

Step III: High-Resolution Supervised Fine-Tuning. The *RetinaLogos* was fine-tuned on higher resolutions, reaching up to 1024×1024, using a dynamic padding strategy from the Next-DIT architecture to enable training with diverse aspect ratios. This allowed the second stage of supervised fine-tuning (SFT) to achieve model convergence more efficiently, even with a relatively limited dataset compared to the scale of natural images. This high-resolution training enhanced image detail, allowing the model to capture finer retinal features.

2.3 Evaluation Standard for Generated Text-to-Retinal Images

As shown in Fig. 2(c), assessing the quality of synthesized retinal images derived from the provided captions is key to their clinical relevance. We employ both downstream task validation and expert evaluation by medical professionals. In particular, in collaboration with ophthalmologists, we designed the first evaluation principles based on five key anatomical structures—the optic disc, macula, retinal vasculature, retinal nerve fibre layer, and pathological lesions—to measure the quality of CFPs generated through a text-driven synthesis approach.

3 Experiments and Discussion

3.1 Dataset and Training Details

Datasets. We evaluate the authenticity of generated CFP using three benchmark datasets: APTOS[12], EyePACs[9] and AIROGS[3]. Additionally, we assess

Table 1. Quantitative Comparison of Synthetic CFP with Real CFP. The FID and KID metrics assess the similarity between generated and real images, while the Inception Score (IS) measures the diversity of generated images. † Results for Lumina-Next are based on weights pre-trained without medical retinal image data. * EyePACs and AIROGS originate from the same institution.

Dataset	FID↓			KID↓			Inception Score
	APTOS	EyePACs	AIROGS	APTOS	EyePACs	AIROGS	
APTOS[12]	-	52.931	42.904	-	0.0417 (0.0015)	0.0342 (0.0015)	1.969
EyePACs[9]	52.931	-	*11.005	0.0417 (0.0016)	-	*0.0066 (0.0008)	2.132
AIRROGS[3]	42.905	*11.005	-	0.0342 (0.0015)	*0.0066 (0.0008)	-	1.993
Average	47.918	52.931	42.904	0.03795	0.0417	0.0342	2.031
MedFusion[16]	77.022	68.162	60.651	0.0871 (0.0016)	0.0748 (0.0013)	0.0716 (0.0015)	1.828
Lumina-Next†[32]	240.406	247.135	251.953	0.1800 (0.0024)	0.1830 (0.0023)	0.1938 (0.0026)	7.615
Ours	**56.078**	**42.437**	**35.190**	0.0369 (0.0012)	0.0230 (0.0008)	0.021 (0.0007)	1.864

Table 2. Performance of Diabetic Retinopathy Grading Classification and Glaucoma Detection. Results for real and synthetic datasets using *RetinaLogos*, with the proposed synthetic CFP data (+ours). Metrics include Accuracy (Acc), F1-Score, and Quadratic Weighted Kappa (QWK).

Training Set	Extra Data	#Samples	Eval Set	Model	Metrics		
					Acc	F1-Score	QWK
IDRiD-Train	N/A	413	IDRiD-Eval	ResNet-50	0.5436	0.4598	0.6223
	N/A	413		ViT-B/16	0.4563	0.3589	0.3848
	+ours	9837+413		ResNet-50	0.6375	0.5183	0.6868
	+ours	9837+413		ViT-B/16	0.5533	0.5400	0.6213
REFUGE2-Train	N/A	640	REFUGE2-Test	ResNet-50	0.8562	0.7842	0.3935
	N/A	640		ViT-B/16	0.9037	0.8615	0.6833
	+ours	9117+640		ResNet-50	0.9375	0.8537	0.5505
	+ours	9117+640		ViT-B/16	0.9762	0.9313	0.8627

the performance of our synthetic data in downstream classification tasks using the IDRiD[21] and REFUGE2[6] datasets. These datasets contain 5,000, 35,126, and 113,893 color fundus images, respectively. The IDRiD dataset includes 516 fundus images. The REFUGE2 dataset consists of 1,200 fundus images, of which we selected 800 labelled images for use in our experiments.

Implementation Details. Our framework is implemented in PyTorch and trained on 8 Nvidia RTX A100 GPUs, we trained *RetinaLogos* with total iterations of 1M with the learning rate of 1×10^{-5}. For the authenticity evaluation, we compared the generated CFPs with real CFPs from the existing open-sourced datasets using metrics of FID, Kernel Inception Distance(KID), and Inception Score(IS). To quantify the text-to-image alignment, we leveraged the existing foundational CLIP-based method [24] to measure the caption-to-image similarity score. In downstream classification tasks, we trained models on different datasets for 100 epochs. ResNet-50 processed images at a resolution of 512×512 pixels, whereas ViT-B/16 operated on 224×224 pixels.

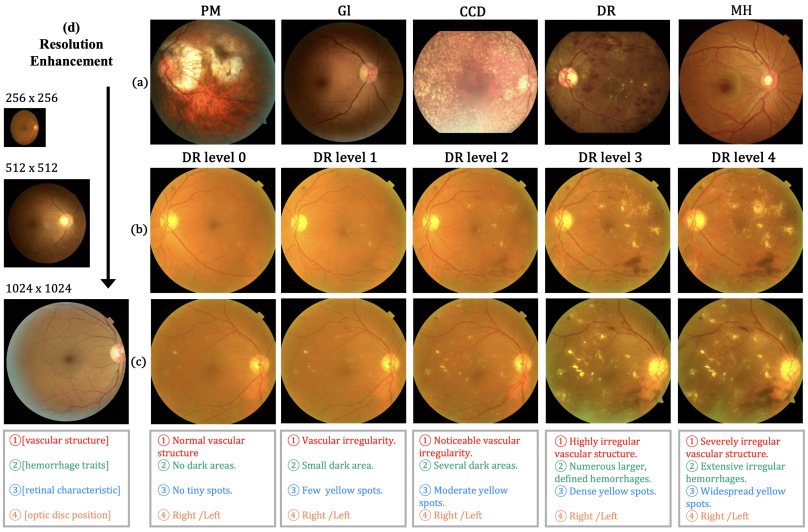


Fig. 3. Visual Comparison of Generated CFPs under Different Stages, Resolutions and Pathological Structures. (a) Disease Categories: Displays different retinal disease types, including pathological myopia (PM), glaucoma (GI), crystalline corneoretinal dystrophy (CCD), diabetic retinopathy (DR), and macular hole (MH). (b) & (c) Diabetic Retinopathy (DR) Levels: Demonstrates the progression of DR across severity levels (0–4). (d) Resolution Enhancement up to 1024 progressively.

3.2 Experimental Results

Comparison Results on Authenticity and Classification Tasks. As shown in Table 1, we compare our method with MedFusion[16] and Lumina-Next[32]. Our method achieves results closest to the real image benchmarks. Visual comparisons are presented in Fig. 3 (a), where CFPs for different eye diseases are generated based on text descriptions. Fig. 3 (b) and (c) demonstrate the controllable generation of Diabetic Retinopathy progression in the left and right eyes under fixed random seeds, achieved by varying the descriptions of anatomical and pathological symptoms. In downstream tasks, we focused on exploring the classification performance of ophthalmological diseases using our augmented training dataset with synthetic CFPs, as detailed in Table 2. Our augmented training data consistently enhances disease classification performance regardless of backbones, leading to an accuracy increase of 5 %–10%. Fig.3 (d) demonstrates the ability of our RetinaLogos to scale up resolution levels, enabling high-quality CFP generation.

Ablation Study. Table 4 summarizes five ablation settings. Exp I (PT) serves as the 256×256 baseline. Exp II introduces prolonged training (PL), but the marginal improvement suggests that simply extending training does not significantly enhance performance. Exp IV applies higher resolution (HR), indicating

Table 3. Clinical Evaluation of Generated Colour Fundus Photographs (CFPs). Authenticity is assessed based on the predictive outcomes for real and synthetic CFPs, including those with no pathological symptoms. The expert evaluation considers the clinical assessment based on clinician-proposed criteria. Values in bold indicate superior performance. The scale of the expert evaluation is scored from 0 to 3 to reflect the level of semantic resemblance between the caption and the generated CFP.

Test Categories	**Evaluation Aspects**	**Prediction Outcome**	
		Real Image	**Synthetic Image**
Authenticity	*Real Image*	**64.29%**	35.71%
	Synthetic Image	**62.07%**	37.98%
Retinal Evaluation	*Optic Disc Structure and Position*	2.11	**2.41**
	Macular Structure and Position	2.10	**2.45**
	Vascular Structure and Position	2.21	**2.55**
	Retinal Nerve Fiber Layer Structure	**2.46**	2.41
	Lesion Structure and Position	1.46	**1.90**
	Overall CFP Image Quality	2.07	**2.28**
	Average	2.06	**2.33**

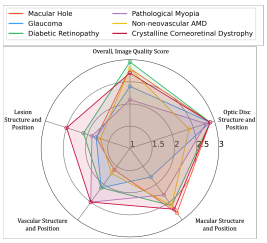


Fig. 4. Comparison of ophthalmologists' evaluation scores on individual retinal disease criteria.

Table 4. Ablation Studies for Generators Configurations and Components. Abbreviations: PT = Pretrain, PL = Prolonged Training, HR = High-Resolution, SR = Clip-Selection & Caption Refinement.

Exp	**Components**				**FID ↓**			**Clip Score ↑**
	PT	PL	HR	SR	APTOS	EyePACs	AIROGS	
O					240.406	247.135	251.953	0.5398
I	✓				63.056	75.386	64.696	0.5485
II	✓	✓			72.389	71.766	63.866	0.5439
III	✓	✓		✓	68.786	73.635	61.465	0.5730
IV	✓	✓	✓		66.525	69.824	61.718	0.5396
V	✓	✓	✓	✓	56.078	42.437	35.190	0.6601

that scaling alone without data refinement offers limited benefit. Exp III incorporates the caption-refinement module (*SR*), highlighting the importance of data quality. Finally, Exp V combines all components, achieving the best overall performance.

Expert Evaluation. To evaluate the generator's ability to produce clinically relevant CFPs, we conducted authenticity tests by comparing the generated images with real ones and performed an expert evaluation focusing on five key aspects. As shown in the Table. 3, 62.07% of the generated CFPs were classified as real, which indicates that the model is capable of producing high-quality images that closely resemble real clinical data. Additionally, Fig. 4 presents an

analysis of the evaluation scores for individual eye diseases based on the generated CFPs.

4 Conclusion

Our work presents *RetinaLogos*, a text-to-image framework that leverages large-scale synthetic retinal caption datasets—comprising 1.4 million entries—to generate high-resolution, clinically relevant retinal images. This approach, which transforms detailed text descriptions into visually rich images capturing key retinal features, has been validated through extensive experiments. Although the controlled generation of the pathological and anatomical structure still leaves room for improvement, particularly in retinal diseases, it shows promising potential to generate CFPs with fine-grained text descriptions in ophthalmology.

Acknowledgments. This work was supported by the National Key R&D Program of China (2022ZD0160101, 2022ZD0160102), the National Natural Science Foundation of China (Grant No.62272450) and Shanghai Artificial Intelligence Laboratory.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Andreini, P., Ciano, G., Bonechi, S., Graziani, C., Lachi, V., Mecocci, A., Sodi, A., Scarselli, F., Bianchini, M.: A two-stage gan for high-resolution retinal image generation and segmentation. Electronics **11**(1), 60 (2021)
2. Bluethgen, C., Chambon, P., Delbrouck, J.B., van der Sluijs, R., Połacin, M., Zambrano Chaves, J.M., Abraham, T.M., Purohit, S., Langlotz, C.P., Chaudhari, A.S.: A vision–language foundation model for the generation of realistic chest x-ray images. Nature Biomedical Engineering pp. 1–13 (2024)
3. De Vente, C., Vermeer, K.A., Jaccard, N., Wang, H., Sun, H., Khader, F., Truhn, D., Aimyshev, T., Zhanibekuly, Y., Le, T.D., et al.: Airogs: Artificial intelligence for robust glaucoma screening challenge. IEEE transactions on medical imaging **43**(1), 542–557 (2023)
4. Demmin, D.L., Silverstein, S.M.: Visual impairment and mental health: unmet needs and treatment options. Clinical ophthalmology pp. 4229–4251 (2020)
5. Du, J., Guo, J., Zhang, W., Yang, S., Liu, H., Li, H., Wang, N.: Ret-clip: A retinal image foundation model pre-trained with clinical diagnostic reports. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 709–719. Springer (2024)
6. Fang, H., Li, F., Wu, J., Fu, H., Sun, X., Son, J., Yu, S., Zhang, M., Yuan, C., Bian, C., et al.: Refuge2 challenge: A treasure trove for multi-dimension analysis and evaluation in glaucoma screening. arXiv preprint arXiv:2202.08994 (2022)
7. Gao, P., Zhuo, L., Lin, Z., Liu, C., Chen, J., Du, R., Xie, E., Luo, X., Qiu, L., Zhang, Y., et al.: Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers. arXiv preprint arXiv:2405.05945 (2024)

8. Go, S., Ji, Y., Park, S.J., Lee, S.: Generation of structurally realistic retinal fundus images with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2335–2344 (2024)
9. Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., et al.: Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. jama **316**(22), 2402–2410 (2016)
10. Islam, J., Zhang, Y.: Gan-based synthetic brain pet image generation. Brain informatics **7**(1), 3 (2020)
11. Jiang, Y., Chen, H., Loew, M., Ko, H.: Covid-19 ct image synthesis with a conditional generative adversarial network. IEEE Journal of Biomedical and Health Informatics **25**(2), 441–452 (2020)
12. Karthik, M., Dane, S.: Aptos 2019 blindness detection. Kaggle https://kaggle.com/competitions/aptos2019-blindness-detection Go to reference in p. 5 (2019)
13. Kim, M., Kim, Y.N., Jang, M., Hwang, J., Kim, H.K., Yoon, S.C., Kim, Y.J., Kim, N.: Synthesizing realistic high-resolution retina image by style-based generative adversarial network and its utilization. Scientific Reports **12**(1), 17307 (2022)
14. Klein, R., Klein, B.E.: Diabetic eye disease. The Lancet **350**(9072), 197–204 (1997)
15. Li, W., Hu, M., Wang, G., Liu, L., Zhou, K., Ning, J., Guo, X., Ge, Z., Gu, L., He, J.: Ophora: A large-scale data-driven text-guided ophthalmic surgical video generation model. arXiv preprint arXiv:2505.07449 (2025)
16. Müller-Franzes, G., Niehues, J.M., Khader, F., Arasteh, S.T., Haarburger, C., Kuhl, C., Wang, T., Han, T., Nolte, T., Nebelung, S., et al.: A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. Scientific Reports **13**(1), 12098 (2023)
17. Ning, J., Marshall, D., Gao, Y., Xing, X., Nan, Y., Fang, Y., Zhang, S., Komorowski, M., Yang, G.: Unpaired translation of chest x-ray images for lung opacity diagnosis via adaptive activation masks and cross-domain alignment. Pattern Recognition Letters **193**, 21–28 (2025)
18. Ning, J., Xing, X., Zhang, S., Ma, X., Yang, G.: Unveiling the capabilities of latent diffusion models for classification of lung diseases in chest x-rays. In: 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2025)
19. Niu, Y., Gu, L., Zhao, Y., Lu, F.: Explainable diabetic retinopathy detection and retinal image generation. IEEE journal of biomedical and health informatics **26**(1), 44–55 (2021)
20. Pham, Q.T., Ahn, S., Shin, J., Song, S.J.: Generating future fundus images for early age-related macular degeneration based on generative adversarial networks. Computer Methods and Programs in Biomedicine **216**, 106648 (2022)
21. Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Meriaudeau, F.: Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. Data **3**(3), 25 (2018)
22. Schmidt-Erfurth, U., Sadeghipour, A., Gerendas, B.S., Waldstein, S.M., Bogunović, H.: Artificial intelligence in retina. Progress in retinal and eye research **67**, 1–29 (2018)
23. Shang, F., Fu, J., Yang, Y., Huang, H., Liu, J., Ma, L.: Synfundus: A synthetic fundus images dataset with millions of samples and multi-disease annotations (2023)
24. 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. Medical Image Analysis **99**, 103357 (2025)

25. Sommer, A., Taylor, H.R., Ravilla, T.D., West, S., Lietman, T.M., Keenan, J.D., Chiang, M.F., Robin, A.L., Mills, R.P., of the American Ophthalmological Society, C., et al.: Challenges of ophthalmic care in the developing world. JAMA ophthalmology **132**(5), 640–644 (2014)
26. Team, G., Riviere, M., Pathak, S., Sessa, P.G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., et al.: Gemma 2: Improving open language models at a practical size. arXiv preprint arXiv:2408.00118 (2024)
27. Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., et al.: Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115 (2024)
28. Yang, Y., Wang, S., Liu, L., Hickman, S., Gilbert, F.J., Schönlieb, C.B., Aviles-Rivero, A.I.: Mammodg: Generalisable deep learning breaks the limits of cross-domain multi-center breast cancer screening. arXiv preprint arXiv:2308.01057 (2023)
29. Zhang, P., Zhao, J., Liu, Q., Liu, X., Li, X., Gao, Y., Li, W.: Fundus image generation and classification of diabetic retinopathy based on convolutional neural network. Electronics **13**(18), 3603 (2024)
30. Zhou, Y., Wang, B., He, X., Cui, S., Shao, L.: Dr-gan: conditional generative adversarial network for fine-grained lesion synthesis on diabetic retinopathy images. IEEE journal of biomedical and health informatics **26**(1), 56–66 (2020)
31. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., et al.: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023)
32. Zhuo, L., Du, R., Xiao, H., Li, Y., Liu, D., Huang, R., Liu, W., Zhao, L., Wang, F.Y., Ma, Z., et al.: Lumina-next: Making lumina-t2x stronger and faster with next-dit. arXiv preprint arXiv:2406.18583 (2024)