npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-026-02804-1

Fundus to fluorescein angiography video generation as a retinal generative foundation model

Check for updates

Weiyi Zhang[1], Jiancheng Yang[2,3], Ruoyu Chen[1], Siyu Huang[4], Pusheng Xu[1], Xiaolan Chen[1], Zhenhan Wang[1], Shanfu Lu[1], Hongyu Cao[5], Xianwen Shang[1], Danli Shi[1,6] ✉ & Mingguang He[1,6,7] ✉

Fundus fluorescein angiography (FFA) is crucial for diagnosing and monitoring retinal vascular issues, but is limited by its invasive nature and restricted accessibility compared to color fundus (CF) imaging. Existing methods for converting CF images to FFA primarily focused on static image generation, often overlooking the dynamic changes in lesions that are critical for comprehensive diagnosis. In this study, we introduce Fundus2Video, an autoregressive generative adversarial network (GAN) model designed to generate dynamic FFA videos from single CF images. Fundus2Video demonstrates exceptional performance in video generation, achieving an FVD of 1497.12 and a PSNR of 11.77, with clinical experts validating the fidelity of the generated videos. Moreover, the model exhibits remarkable transferability across ten external public datasets, successfully addressing tasks including blood vessel segmentation, retinal disease diagnosis, systemic disease prediction, and multimodal retrieval, showcasing impressive zero-shot and few-shot capabilities. These findings position Fundus2Video as a robust, non-invasive alternative to traditional FFA examinations and a versatile retinal generative foundation model. By capturing both static and temporal retinal features, Fundus2Video provides transferable representations for multiple ophthalmic downstream tasks and may motivate analogous cross-modality studies in other medical imaging domains.

Color fundus photography (CF) and fundus fluorescein angiography (FFA) are vital diagnostic tools for retinal diseases. CF is favored for its simplicity and affordability, particularly in resource-limited settings like rural and remote areas. However, CF has limitations in detecting lesions like macular edema, retinal non-perfusion, and neovascularization, essential for diagnosing vascular diseases like diabetic retinopathy (DR) and guiding treatment. In contrast, FFA excels in identifying and monitoring these lesions by using an injected dye to dynamically highlight pathological changes. Despite its diagnostic benefits, FFA is invasive and carries risks of potential side effects, including nausea and anaphylactic shock[1]. Therefore, developing non-invasive, safe, and cost-effective alternatives is crucial.

Generative adversarial networks (GANs) have shown promise in translating CF images into realistic FFA images, as demonstrated by the Fundus2FA model[2]. This method can robustly translate CF into realistic FFA images and improve DR stratification. This line of work has been extended in ophthalmology to related cross-modality angiographic generation settings, further supporting the feasibility of recovering angiographic information from non-invasive fundus imaging[3–5]. However, existing methods are limited to generating static snapshots, lacking the continuous and dynamic vascular changes observed in real FFA[6]. In broader medical imaging research, autoregressive GAN architectures have been explored for sequential synthesis and temporally coherent generation, such as in cardiac MRI reconstruction using temporally aware volumetric GANs[7], or physiological video generation in MRI using latent-space guided GANs[8]. More recently, text-to-video generation has been applied to FFA video synthesis, highlighting growing interest in dynamic angiographic generation[9]. Despite these advances, prior work has not directly addressed the generation of patient-specific FFA video sequences from color fundus photographs, leaving a notable gap in modeling retinal flow dynamics from routine fundus imaging[10,11].

Recent advancements in deep learning have led to the development of foundation models capable of performing general tasks[12]. These models are pre-trained on large datasets to extract generalized feature representations,

[1]School of Optometry, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China. [2]ELLIS Institute Finland, Espoo, Finland. [3]Aalto University, Espoo, Finland. [4]School of Computing, Clemson University, Clemson, SC, USA. [5]University of Science and Technology of China, Anhui, China. [6]Research Centre for SHARP Vision (RSCV), The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China. [7]InnoHK, Centre for Eye and Vision Research (CEVR), Hong Kong Science Park, Hong Kong SAR, China. ✉e-mail: danli.shi@polyu.edu.hk; mingguang.he@polyu.edu.hk

which can be easily adapted to specific tasks by fine-tuning with small amounts of labeled data. In ophthalmology, foundation models based on the masked autoencoder[13–15] and the contrastive language-image pre-training framework[15–17] have recently been further extended to retinal language-image modeling and multimodal retinal analysis. Despite their success, these models require large datasets, extensive training time, and significant computational resources. Furthermore, while segmentation-oriented general medical models such as Medical-SAM have broadened the scope of foundation models in medical imaging, evidence remains limited that a lightweight cross-modality generative model can generalize across diverse downstream retinal tasks, including segmentation, multiple classification settings, and multimodal retrieval[18,19].

To address the challenges inherent in dynamic FFA generation and to harness the potential of generative models as domain-specific foundation models, we propose Fundus2Video, a novel framework that generates FFA videos from static CF images. This cross-modality generative model excels at video generation and can be adapted as a lightweight pre-trained representation model for various tasks, including ophthalmic and systemic disease classification, segmentation, and image retrieval across ten multinational datasets, under zero-shot, few-shot, and supervised experimental settings. Fundus2Video supports the potential of lightweight modality translation frameworks as task-specialized generative foundation models for retinal imaging.

The remainder of this paper is organized as follows. The Results section presents the quantitative and qualitative outcomes of Fundus2Video's video generation, alongside its performance on diverse downstream tasks, including vessel segmentation, disease diagnosis, and multimodal retrieval. The Discussion section interprets these findings, explores the clinical implications, acknowledges the limitations of our study, and suggests directions for future research. Finally, the Methods section provides a detailed description of the datasets, the Fundus2Video architecture, the training procedures, and the evaluation metrics employed.

Results

Figure 1 illustrates the study design. Fundus2Video was developed using 1,956 CF images paired with their corresponding FFA videos. These videos encompass frames from the entire FFA examination process, including the arterial, venous, and late phases, totaling 72,851 FFA frames. The dataset includes various lesions such as macular edema, microaneurysms, capillary leakage, and retinal non-perfusion. Fundus2Video employs an autoregressive GAN architecture, enhanced by an unsupervised clinical knowledge mask (Fig. 1a). To demonstrate its versatility as a retinal foundation model, we evaluated Fundus2Video across ten distinct datasets, covering a wide range of tasks (Fig. 1b).

Quality of generated FFA videos

Model comparisons. We evaluated Fundus2Video against existing image-to-video translation methods, including the auto-encoder-based Seg2Vid[20], diffusion model-based Med-ddpm[21], ConsistI2V[22], and streamingSVD[23]. Each model generates a 12-frame video covering the entire FFA examination process, including the arterial, venous, and late phases. Table 1 shows that our model outperforms all others across various objective quality metrics, achieving the lowest Fréchet Video Distance (FVD) of 1497.12 and Learned Perceptual Image Patch Similarity (LPIPS) of 0.2131, along with the highest Structural Similarity Index (SSIM) of 0.3795 and Peak Signal-to-Noise Ratio (PSNR) of 11.77. Beyond quality improvements, Fundus2Video also demonstrates a favorable efficiency–performance trade-off. Compared with diffusion-based methods, it requires substantially shorter training time, significantly faster inference, and markedly lower computational cost in terms of TFLOPs, while still delivering superior video quality.

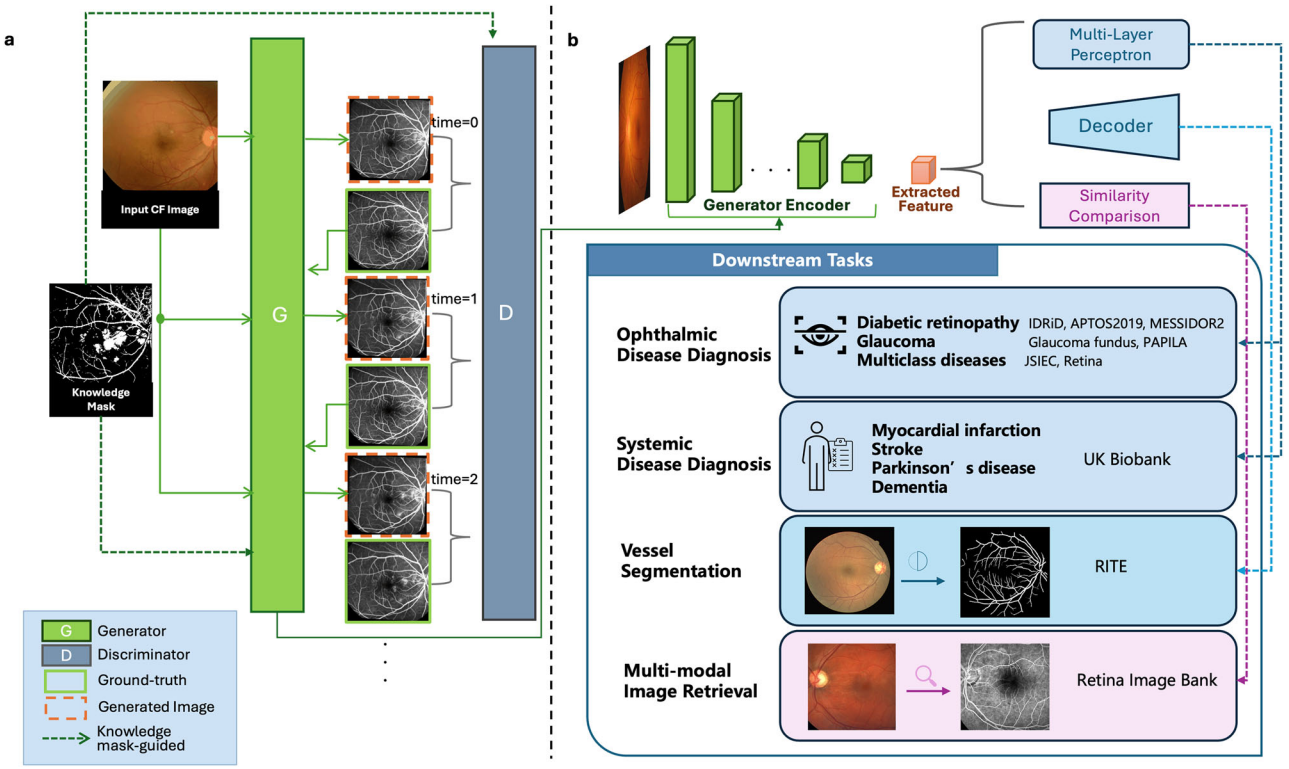


Fig. 1 | Study overview. a Architecture of the Fundus2Video Model: The model is based on an autoregressive GAN. The generator produces one frame at a time, using the output from the previous time step and the CF image embedded in different channels as input. During training, the network is guided by an unsupervised knowledge mask, created by calculating the difference between the first and last frames. This approach leverages clinical insights that more severe lesions result in more significant fluorescent changes. The knowledge mask effectively highlights pathological regions in the FFA frames. **b** Adapting Fundus2Video for Various Downstream Tasks: the generator encoder from Fundus2Video was used to accomplish various downstream tasks, including ophthalmic disease diagnosis, systemic disease prediction, vessel segmentation, and multi-modal image retrieval.

Table 1 | Comparison of video generation methods

Models	FVD ↓	SSIM ↑	PSNR ↑	LPIPS ↓	Training time (hrs)	Inference time (sec / video)	TFLOPs
Seg2Vid	2301.23	0.2871	10.54	0.2513	~ 4	~ 0.8	~ 0.8– 1.2
Med-ddpm	2401.43	0.2552	10.73	0.2407	~ 120	~ 18	~ 1.5– 2.5
ConsistI2V	2231.35	0.2598	10.84	0.2399	~ 80	~ 6	~15–30
streamingSVD	2157.32	0.2893	10.97	0.2461	~ 150	~ 12	~25– 50
Fundus2Video	**1497.12**	**0.3795**	**11.77**	**0.2131**	~ 28	~ 0.9	~1.0– 1.8

Fundus2Video outperforms other models across multiple video quality metrics, including FVD, SSIM, PSNR, and LPIPS. In addition, Fundus2Video achieves competitive training efficiency, fast inference speed, and low computational cost, demonstrating clear advantages over the auto-encoder-based Seg2Vid and diffusion-based methods such as Med-ddpm, ConsistI2V, and streamingSVD under a unified evaluation setting.

Visualization results. Visualization examples are shown in Fig. 2. Fundus2Video produces realistic FFA videos regarding retinal structures and lesional areas, such as microaneurysms, leakage, nonperfusion area, laser scar, etc.

Human assessment. Two ophthalmologists reviewed the results from all methods in Table 1 and concluded that Fundus2Video significantly outperformed the others. They then conducted a quality assessment on 50 randomly selected FFA videos generated by Fundus2Video from the test set. The evaluation focused on vascular perfusion, lesion dynamics, and overall coherence, comparing the generated videos to corresponding CF images and ground-truth FFA videos. This evaluation was specifically designed to account for vascular flow dynamics in the absence of pixel-level ground truth, ensuring that temporal perfusion and lesion changes were clinically validated. Scores ranged from 1 (excellent quality) to 5 (very poor quality). The scoring rules and results are summarized in Table 2. The experts gave subjective quality scores of 1.28 and 1.34, respectively. The inter-rater agreement was almost perfect ($\kappa = 0.90$, $P < 0.001$), underscoring the reliability and consistency of the subjective quality assessment. To further assess robustness and mitigate potential overfitting to a specific dataset, we conducted an additional human evaluation using external datasets collected from diverse populations and imaging devices. Specifically, 20 CF images (4 per dataset) were randomly sampled from IDRID[24] (India, Kowa VX-10 alpha), MESSIDOR2 (France, Topcon TRC-NW6), PAPILA[25] (Spain, Topcon TRC-NW400), JSIEC[26] (China, ZEISS FF450 Plus IR and Topcon TRC-50DX), and UK Biobank[27] (United Kingdom, Topcon 3DOCT-1000-Mk2). Using the same evaluation protocol, the two experts assigned mean quality scores of 1.10 and 1.25, respectively, indicating that Fundus2Video maintains high visual and clinical fidelity across diverse external datasets and imaging conditions.

Downstream performance in Zero-shot and Few-shot settings for vessel segmentation

Fundus2Video's generator exhibits strong zero-shot and few-shot segmentation capabilities. We validated this on the vascular segmentation dataset RITE[28], conducting few-shot experiments on retinal vessels with support numbers of 0, 1, 2, 4, 8, and 16. We compared Fundus2Video with Fundus2FA, a GAN model based on Pix2PixHD[29] generating venous and late-phase single frame FFA (here we used the venous-phase model weight), and the recent segmentation foundation model Medical-SAM2[30], known for its medical generalization capability. Detailed numerical comparisons and visual results are shown in Fig. 3.

As illustrated in the bar charts of Fig. 3a, Fundus2Video outperformed all models in all settings in terms of the area under the receiver operator characteristic curve (AUROC) and Dice scores, particularly compared to Medical-SAM2 and Pix2pixHD. Notably, Fundus2Video achieves an AUROC of 0.943 in the zero-shot setting and a Dice score of 0.733 in the one-shot setting, showcasing its powerful few-shot segmentation capabilities. The visual comparison of one-shot segmentation results in Fig. 3b further emphasizes Fundus2Video's superiority, where other methods display some loss of detail in vascular segmentation, while Fundus2Video's results closely resemble the ground truth with remarkably rich details. This demonstrates Fundus2Video's profound understanding and cross-modal reconstruction ability for fundus images, particularly in capturing fine-grained vascular structures.

Downstream performance in zero-shot, few-shot, and supervised training for ophthalmic disease diagnosis

To evaluate Fundus2Video's potential as a foundation model for diagnosing various ophthalmic diseases from CF images, we evaluated it on seven public datasets. We compared Fundus2Video with Fundus2FA and RETFound, the previous state-of-the-art foundation model in ophthalmology. To ensure statistical robustness, we performed a repeated random sub-sampling validation (5 iterations). The reported results are the average over these five independent experiments, each with a different random split of the dataset.

In the zero-shot classification, where the encoder from the models was used without further training, Fundus2Video consistently outperformed others across all datasets. Notably, on glaucoma datasets such as Glaucoma Fundus[31] and the multi-disease dataset JSIEC[26], Fundus2Video achieved the area under the receiver operator characteristic curve (AUROC) scores more than 30% higher than RETFound. This demonstrates Fundus2Video's strong ability to understand features from external fundus images across diverse sources. Results are presented in Fig. 4a and Supplementary Table 1.

We further evaluated the models' performance in few-shot learning settings with 1, 2, 4, and 8 support examples per class. Fundus2Video showed a clear advantage in the one-shot setting across all seven datasets, indicating strong feature generalization and learning capabilities even with minimal labeled data. The results suggest that Fundus2Video effectively captures essential retinal features with limited supervision. Detailed results are shown in Fig. 5a and Supplementary Table 2.

Finally, we fine-tuned and tested the models by splitting the datasets into 55:15:30% (train:val: test). The corresponding results are presented in Fig. 6a and Supplementary Table 3. In diabetic retinopathy datasets (MESSIDOR2, IDRID[24], and APTOS2019), Fundus2Video improved AUROC by up to 1.5% (e.g., MESSIDOR2) and increased area under the precision-recall curve (AUPR) scores, demonstrating better classification of disease stages compared to using static CF images alone. For glaucoma-focused datasets (PAPILA[25] and Glaucoma Fundus[31]), Fundus2Video achieved AUROC improvements of up to 0.8% and maintained high AUPR values, indicating enhanced sensitivity and specificity in detecting glaucoma-related changes. Across multi-class eye disease datasets (JSIEC[26] and Retina), Fundus2Video consistently raised AUROC by up to 0.6% and improved AUPR, with statistically significant improvements (p-value < 0.001). These findings underscore Fundus2Video's efficacy in enhancing overall classification accuracy across diverse eye conditions, emphasizing the utility of dynamic FFA data in improving diagnostic precision.

Downstream performance in zero-shot, few-shot, and supervised training for systemic disease diagnosis

To evaluate Fundus2Video's improvement in systemic disease diagnosis using CF images, we compared it with Fundus2FA and RETFound across Stroke, Dementia, Parkinson's Disease (PD), and Myocardial Infarction (MI) using the UK Biobank dataset[27]. Similar to our ophthalmic disease

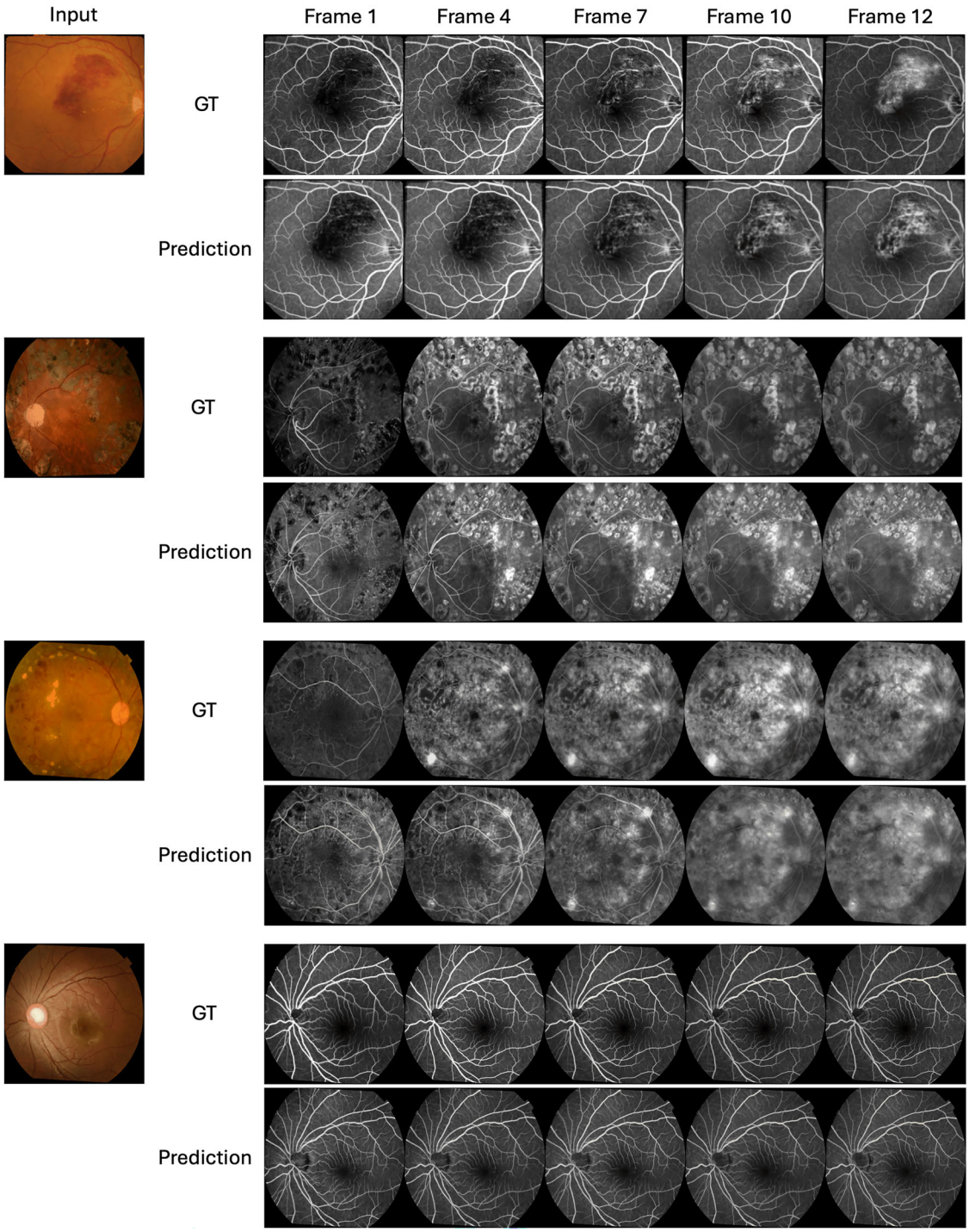


Fig. 2 | Visualization examples of generated videos. GT = Ground Truth. The original generated videos consist of 12 frames, with 5 frames sampled for display. The first case includes lesions of leakage and nonperfusion areas. The second and third cases feature microaneurysms, leakage, nonperfusion areas, and laser scars. The fourth case shows no abnormalities in FFA. Fundus2Video effectively generates smooth FFA series that accurately replicate anatomical and pathological features. The frames are displayed without an absolute fluorescence intensity scale in accordance with clinical convention, as interpretation of FFA relies on relative temporal changes in fluorescence rather than absolute intensity values.

tests, we evaluated the models in zero-shot, few-shot (support Number=1, 2, 4, 8, 16), and supervised learning settings. Results are shown in Figs. 4–6b and Supplementary Table 4-6. Overall, Fundus2Video showed favorable AUROC and AUPR performance compared with Fundus2FA and RETFound across the systemic disease prediction tasks.

Notably, in the zero-shot setting, Fundus2Video outperformed RETFound by over ten percentage points in both AUROC and AUPR when predicting Dementia and PD. In the one-shot setting, the performance gap was even larger for Parkinson's disease. These findings suggest that the learned representations capture retinal features relevant to vascular and neurodegenerative systemic disease prediction.

Downstream performance in multi-modal image retrieval

To further explore Fundus2Video's robust feature representation beyond CF and FFA, we investigated its performance in image retrieval across multiple modalities. Accurate image retrieval in clinical practice aids in case

Table 2 | Human visual assessment results. A. Results. B. Assessment criteria

A. Results.				
Test set	**Score by Rater 1**	**Score by Rater 2**	**Kappa value**	***P*-value (95% confidence level)**
Internal Test	1.28	1.34	0.9	<0.001
External Test	1.1	1.25	0.63	0.47
B. Assessment criteria.				
Scale	**Quality**	**Description**		
1	Excellent	Similar morphology and fluorescent changes of retinal vessels and lesions		
2	Good	Slight morphology and fluorescent changes of retinal vessels and lesions		
3	Normal	Moderate variation in morphology and fluorescent changes of retinal vessels and lesions		
		Indicates lesions of diagnostic significance		
4	Poor	Large variation in morphology and intensity of retinal vessels and lesions		
		Without lesions of diagnostic significance		
5	Very Poor	Large variation in morphology and intensity of choroidal and retinal vessels		

The two aspects of the scoring criteria are vascular perfusion and conformity of lesion dynamics to clinical patterns in the video. Scores ranged from 1 (excellent quality) to 5 (very poor quality).

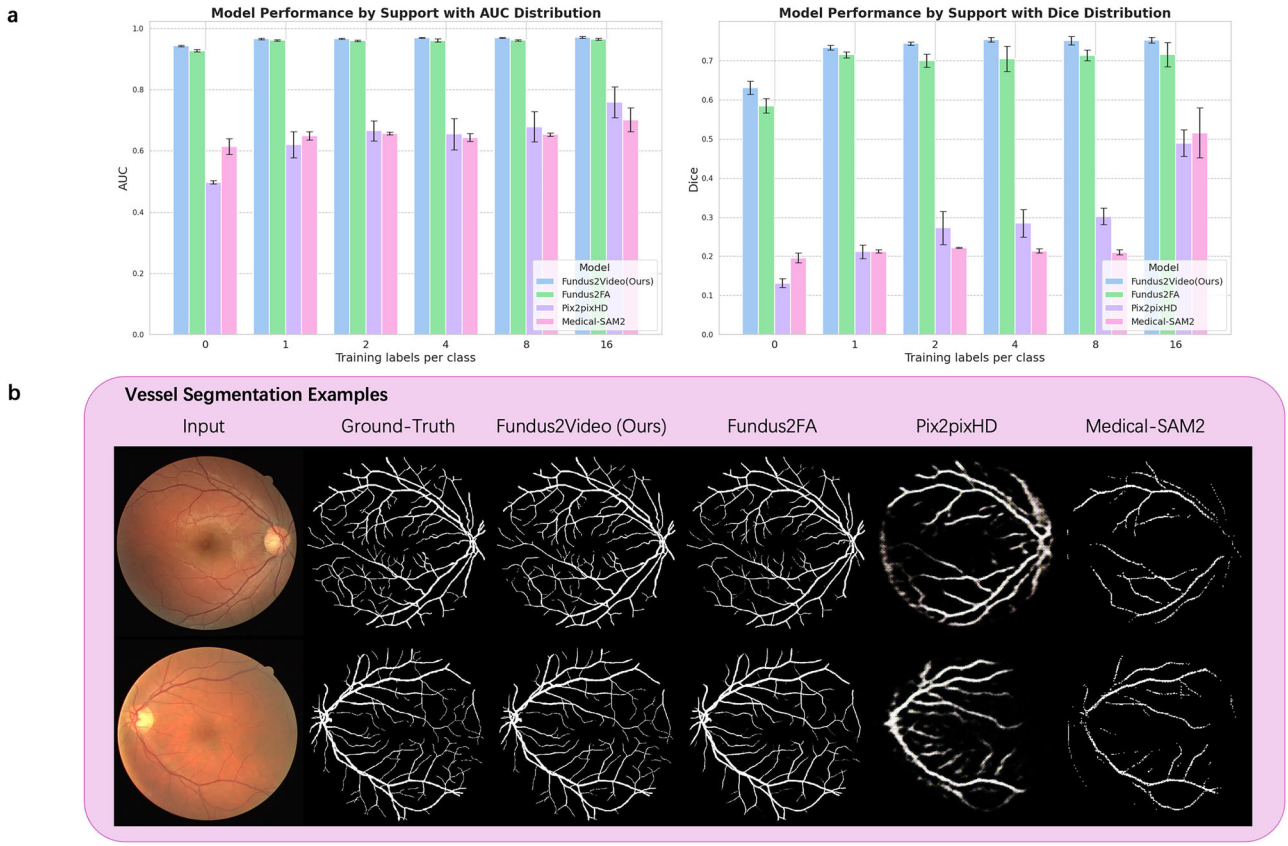


Fig. 3 | Vessel segmentation results. a Model performance comparison. AUROC = area under the receiver operator characteristic curve. We compared the segmentation capabilities of four models by setting the number of training samples to 0, 1, 2, 4, 8, and 16. Five experiments were conducted for each setting, with different datasets randomly sampled and divided. Fundus2Video achieved the best AUROC and Dice scores across all settings, significantly outperforming Pix2pixHD and Medical-SAM2. **b** Vessel segmentation visualization results of the one-shot setting. The output of Fundus2Video closely approximates the ground truth with rich details, whereas the results from other models are visibly incomplete.

comparison, disease progression tracking, and identifying rare conditions, especially with multimodal imaging data. This is particularly useful for searching similar cases in large-scale medical databases or retrieving specific imaging patterns for diagnostic support.

We compared the multimodal image retrieval abilities of Fundus2Video, Fundus2FA, and RETFound. Validation was performed on two external multimodal datasets: AngioReport[32] and Retina Image Bank[33]. The AngioReport dataset includes images from indocyanine green angiography (ICGA) and FFA, while the Retina Image Bank encompasses 14 ophthalmic imaging modalities, including rare diseases. Retrieval performance and visualized examples are shown in Fig. 7 and Supplementary Table 7.

Across both datasets, Fundus2Video achieved the best performance in terms of top-1, top-5, top-10, and average recall. In the AngioReport dataset, Fundus2Video's average recall exceeded 0.8. As shown in the retrieval examples in Fig. 7b, Fundus2Video not only successfully performed cross-

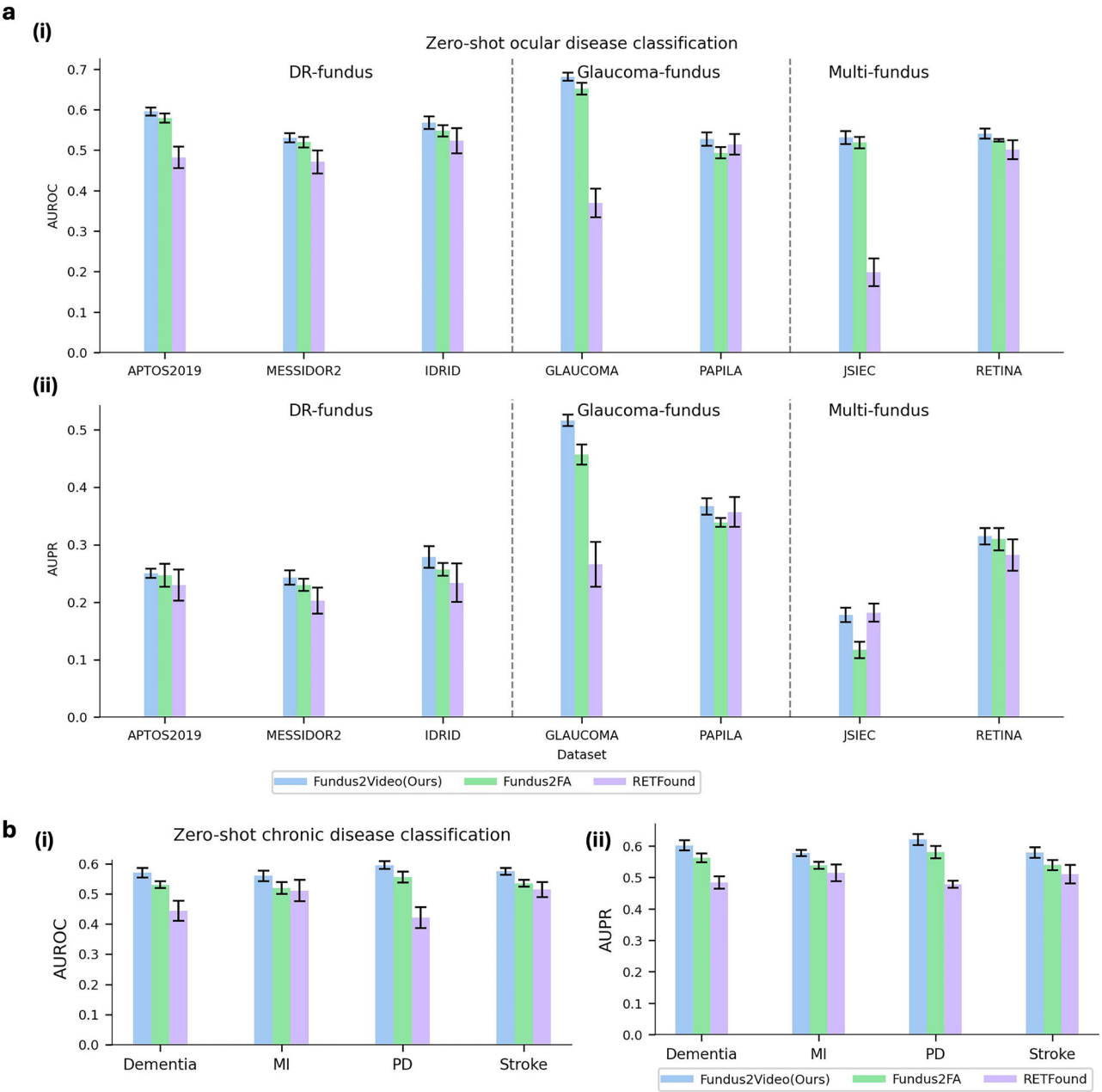


Fig. 4 | Zero-shot disease classification results. AUROC = area under the receiver operator characteristic curve, AUPR = area under the precision-recall curve, MI = Myocardial Infarction, PD = Parkinson's Disease. **a** Results of ocular disease classification. Fig a.(i) shows the AUROC results, and Fig a.(ii) shows the AUPR results. Experiments on each dataset were averaged over five runs. Fundus2Video achieved a higher mean AUROC compared to the other two models, Fundus2FA and RETFound, across all datasets ($P < 0.001$). The P-value was calculated by comparing the prediction results of Fundus2Video with the prediction from the other model with a higher AUROC result using a t-distribution. **b** Results of chronic disease classification. Fig b.(i) shows the AUROC results, and Fig b.(ii) shows the AUPR results. Experiments on each dataset were averaged over five runs. Fundus2Video demonstrated significantly better AUROC and AUPR across all tasks, with all P-values < 0.001.

modality retrieval for the same patient (Case 1), but also enabled retrieval of images from the same modality and patient with different viewing angles (Case 2). These results highlight Fundus2Video's exceptional ability to handle multimodal images.

Discussion

Multi-modality examinations provide complementary information, and Fundus2Video addresses part of this gap by generating non-invasive FFA videos from static CF images. Beyond visual authenticity validated by retinal specialists, we demonstrated its potential as a retinal foundation model across multiple downstream tasks, including vessel segmentation, retinal disease diagnosis, systemic disease prediction, and multimodal retrieval. Fundus2Video captures both cross-modal static and temporal retinal features, enabling the representation of complex inter-modality relationships. We are acutely aware that the hemodynamic realism of the generated videos is critical to their potential clinical value; failure to accurately reflect individual variations in retinal circulation and true pathophysiology may pose a risk of misinterpretation due to artificially synthesized dynamic patterns. In this study, we ensured the clinical coherence of the generated sequences through quantitative metrics and extensive subjective evaluations by retinal specialists. Validation was further performed on diverse external datasets, with a focus on vascular perfusion and lesion dynamics.

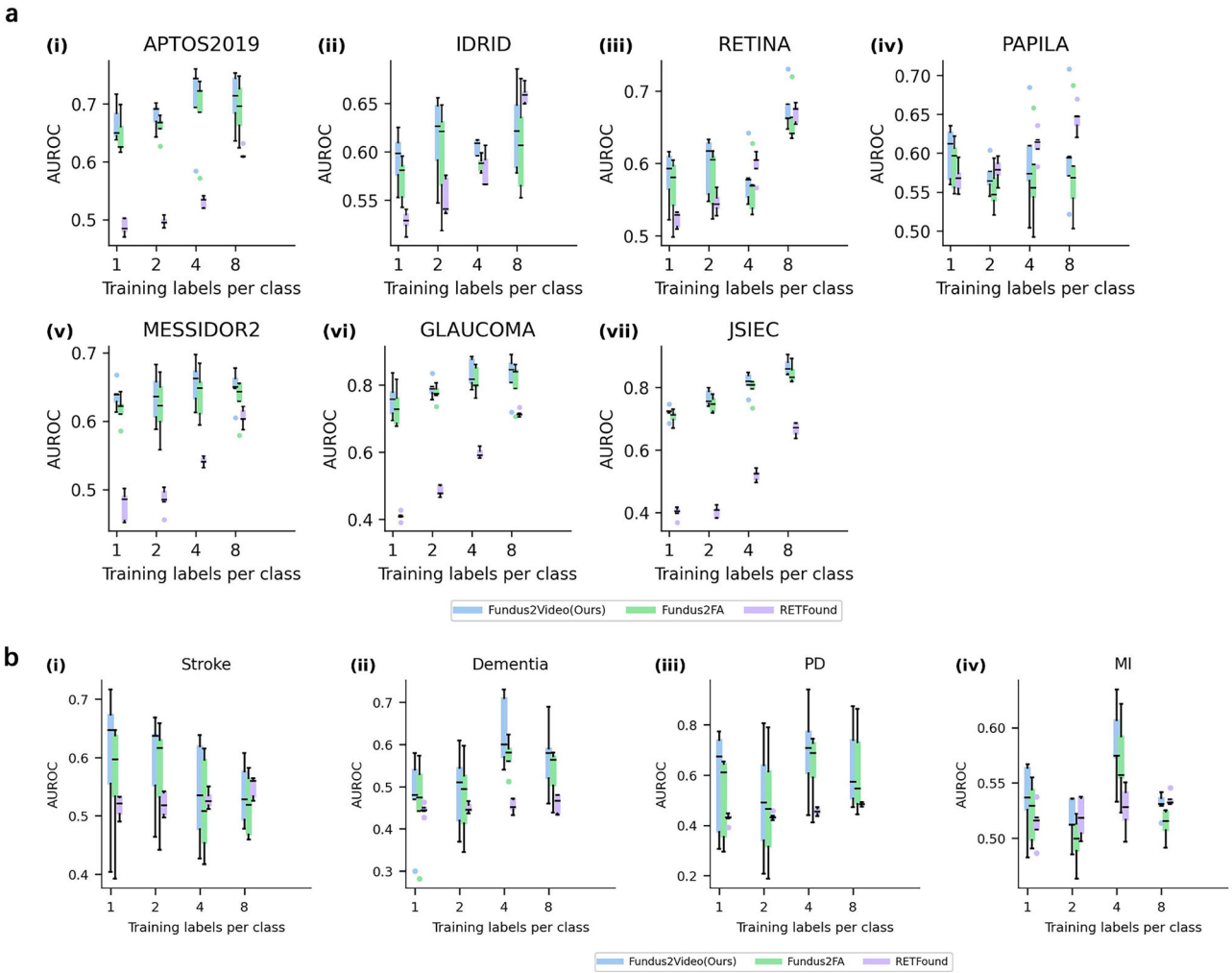


Fig. 5 | Few-shot classification results. AUROC = area under the receiver operator characteristic curve. We compared the few-shot learning capabilities of different models across various datasets, considering multiple support numbers for each class. For each support number, we sampled five different sets of training examples and conducted five times of experiments. The boxes represent the quartiles, and the whiskers extend to data points within 1.5 times the interquartile range. AUPR results are shown in Supplementary Fig. 1. **a** Ophthalmic diseases. We selected support numbers of 1, 2, 4, and 8. Notably, when the support number is 1, Fundus2Video significantly outperforms the other two models across all datasets ($P < 0.001$), demonstrating its exceptional generalization capability. Additionally, on the APTOS2019, MESSIDOR2, Glaucoma Fundus, and JSIEC datasets, Fundus2Video consistently outperforms the other two models across all support numbers ($P < 0.001$). **b** Chronic diseases. We evaluated few-shot classification with support numbers of 1, 2, 4, 8, and 16 per class. Fundus2Video showed consistently strong performance across systemic disease tasks, with the clearest advantage in dementia, Parkinson's disease, and myocardial infarction. For stroke, the advantage was more evident at lower support numbers.

A strength of Fundus2Video is its strong performance in classifying ophthalmic diseases like DR and glaucoma. DR involves microvascular changes such as capillary damage and neovascularization, best viewed by FFA[34]. In contrast, glaucoma causes optic nerve damage and progresses silently until significant vision loss occurs[35]. Fundus2Video enhances detection by learning structures like retinal vessels, optic disks, and lesions that are evident on FFA, surpassing static fundus image representations[13]. Furthermore, to address potential concerns regarding overfitting and generalizability, an external validation was performed using datasets from five different countries and various imaging devices. The expert assessment suggested that Fundus2Video preserved good visual quality and clinically plausible lesion dynamics across these diverse sources. These results support the potential of this approach for future screening-oriented research, although prospective clinical validation is still required.

Moreover, Fundus2Video excels in retinal vessel segmentation, particularly in few-shot and zero-shot learning settings. In particular, the model achieved an AUROC of 0.965 in the one-shot setting, suggesting that the pretrained encoder captures informative vascular features relevant to retinal vessel delineation. This is consistent with our previous findings that cross-modality pretraining can benefit vessel segmentation[36]. The fine-grained vascular information learned by Fundus2Video may also contribute to the observed gains in both ophthalmic and systemic disease prediction tasks.

Another contribution of Fundus2Video is to demonstrate that generative models can be lightweight, task-specific foundation models. Unlike traditional, larger models, Fundus2Video efficiently learns from CF and FFA modalities with fewer computational demands. Its generator encoder has about 60 million parameters, compared to the 307 million in Vision Transformer Large (ViT-Large) used in RETFound[13,37]. This efficiency enables high performance across diverse tasks, including segmentation, without the burdens of larger architectures.

Discriminative and generative foundation models each have unique strengths[38]. Discriminative models are ideal for classification tasks and perform well with large labeled datasets, valued for their simplicity and effectiveness. On the other hand, generative models are used for self-reconstruction in self-supervised learning or creating synthetic datasets for data augmentation, which is beneficial in the medical field where labeling is

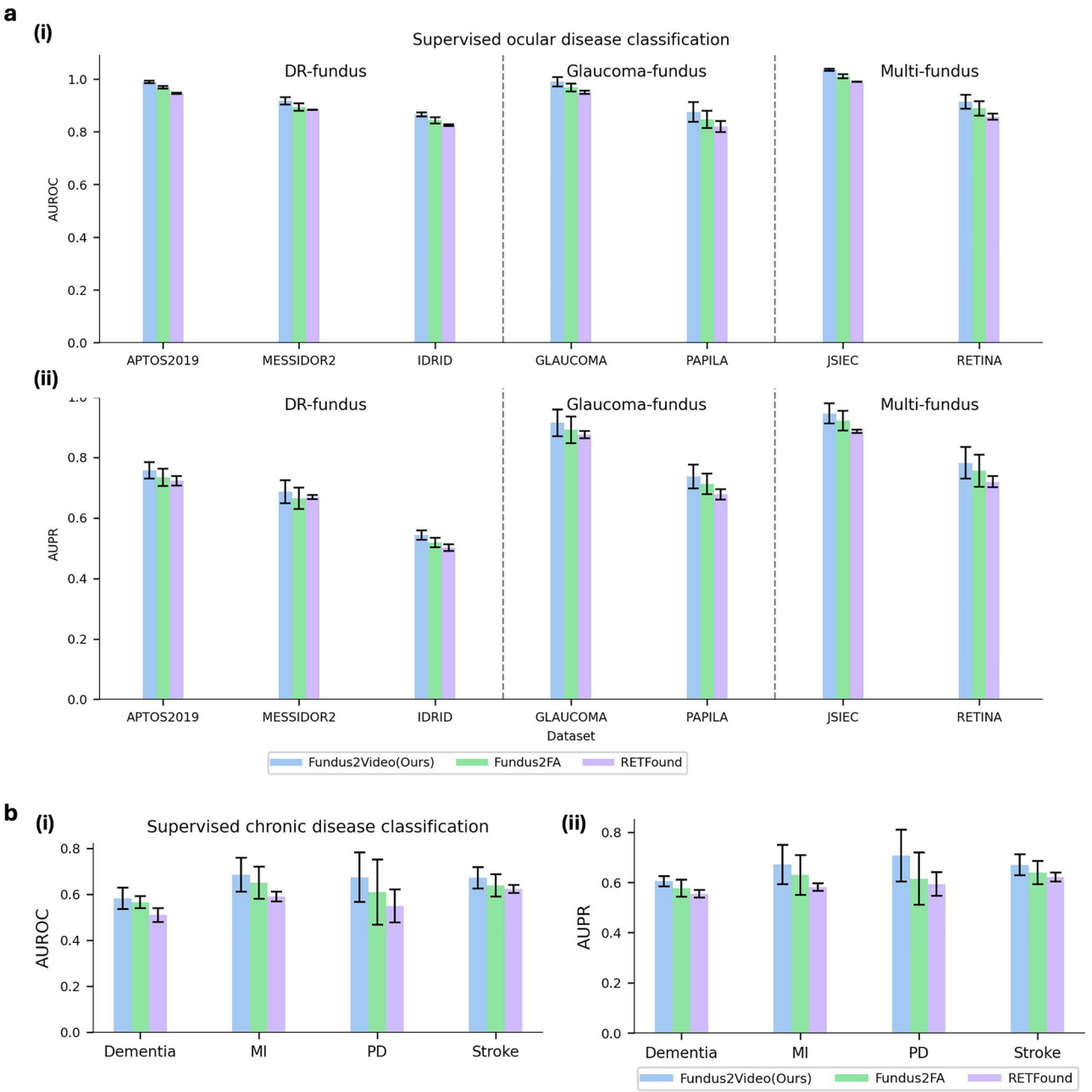


Fig. 6 | Supervised classification results. a Results of ocular disease classification. Fig a.(i) presents the AUROC results, while Fig a.(ii) displays the AUPR results. AUROC = area under the receiver operator characteristic curve, AUPR = area under the precision-recall curve. Each dataset's experiments were averaged over five runs. Fundus2Video achieved a higher mean AUROC compared to the other two models, Fundus2FA and RETFound, across all datasets. Regarding AUPR, Fundus2Video outperformed the other two models except for the MESSIDOR2 dataset. **b** Results of chronic disease classification. Fig b.(i) shows the AUROC results, and Fig b.(ii) shows the AUPR results. Each dataset's experiments were averaged over five runs. For AUROC, Fundus2Video was the best performer across all tasks except for Dementia, which was slightly lower than Fundus2FA ($P > 0.05$). Fundus2Video also demonstrated superior AUPR across all tasks.

costly and time-consuming. Fundus2Video bridges these approaches by combining generative cross-modality learning with downstream representation transfer. In this setting, paired clinical imaging data can be used more effectively with limited task-specific labeling, and the results suggest that generative pretraining can support multiple downstream retinal tasks. This framework may also motivate exploration of analogous cross-modality generative approaches in other medical imaging domains.

While Fundus2Video demonstrates promising performance, several limitations warrant consideration. As noted earlier, the model's zero-shot capability, while powerful, may not capture the most subtle pathological nuances for complex diseases, necessitating fine-tuning for expert-level precision. The model remains partially a black box, making it an ongoing challenge to quantify the physiological fidelity of generated videos, particularly for subtle vascular dynamics like microaneurysm leakage or slow perfusion changes. Although our model successfully captures certain leakage patterns, phenomena requiring extended temporal observation remain difficult to fully represent with the current 12-frame sequences. Furthermore, the training dataset of 1956 CF images with corresponding FFA frames, while sufficient to demonstrate feasibility for this specific task, is relatively modest compared to large-scale foundation models. Although the model is lightweight and performed competitively against larger baselines, the relatively modest training set remains an important limitation when interpreting the scope of generalizability. Additionally, while external validation across diverse populations and imaging devices confirms some

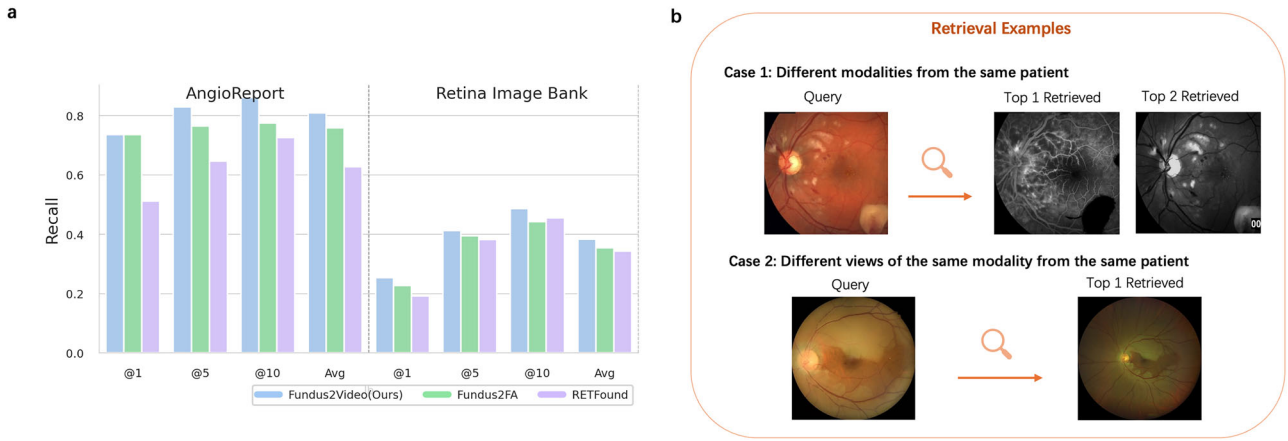


Fig. 7 | Zero-shot multi-modal image retrieval results. a Comparison results of different models. Similarity in the embedding space was computed between the query image and all the other images in the test set. The top K most similar images were retrieved, where K = 1, 5, 10. We report Recall@K and the mean recall. Fundus2Video outperforms all other models across both datasets. **b** Retrieval examples using Fundus2Video. Case 1 demonstrates the model's ability to retrieve images of different modalities from the same patient, while Case 2 shows the model's capability to retrieve images of different views within the same modality from the same patient.

generalizability, the coverage of rare diseases and camera types remains incomplete. The clinical applicability of the generated videos requires further validation through large-scale, multi-center prospective studies involving real patient data from diverse clinical settings to confirm diagnostic reliability and evaluate clinical utility. Any future clinical deployment must be accompanied by clear disclosure of the AI-generated nature of these videos and appropriate clinician training. From a technical perspective, while the current framework focuses on pixel-aligned FFA generation, incorporating additional modalities may further improve representation quality and downstream task performance[3,29]. Future work will address these limitations by extending sequence length, refining temporal modeling, scaling to larger and more diverse datasets through multi-center collaborations, and developing more flexible frameworks capable of handling diverse data types through methods such as contrastive learning.

Several technical challenges are crucial for the real-world deployment of Fundus2Video, including modality alignment, feature fusion, domain adaptation, and computational scalability. In our current framework, the autoregressive GAN architecture, regularized by spatial constraints (e.g., the knowledge mask) and temporal consistency losses, encourages the learning of a coherent cross-modal latent space, thereby partially mitigating the difficulty of aligning structural (CF) and functional (FFA) representations. Furthermore, the lightweight GAN design ensures efficient feature encoding and retrieval, offering advantages in scalability and potential deployment on resource-constrained devices compared to large transformer-based models. Despite these considerations, robust generalization across unseen domains and highly scalable multimodal retrieval remain open challenges. Future work will focus on developing more advanced domain adaptation techniques, such as adversarial feature alignment or domain-specific attention mechanisms, to improve robustness to variations in imaging devices and patient populations. Moreover, we plan to explore scalable multimodal pretraining strategies, potentially based on contrastive learning, to foster tighter integration of features from CF, FFA, and other clinically relevant data sources such as diagnostic reports.

In conclusion, this study presents Fundus2Video, an innovative cross-modality generative model that transforms static CF images into dynamic FFA videos and yields transferable representations for multiple downstream tasks. By learning from both CF and FFA modalities, Fundus2Video enhances diagnostic accuracy for various retinal diseases and shows significant potential for systemic disease screening. These findings support the potential of lightweight generative pretraining in ophthalmic imaging, while further multi-center and prospective validation remains necessary.

Methods

Study Population

Data for algorithm development. The dataset used for developing Fundus2Video comprises 1956 CF images and 72,851 corresponding FFA frames obtained from 1360 anonymized patients who underwent FFA examinations between 2016 and 2019 in China[2]. The CF images were captured using Topcon TRC-50XF and Zeiss FF450 Plus cameras, with resolutions ranging from 1110 × 1467 to 2600 × 3200 pixels. Meanwhile, FFA images were acquired using Zeiss FF450 Plus and Heidelberg Spectralis cameras, featuring a resolution of 768 × 768 pixels. During training, the ground-truth FFA videos also consist of 12 frames, with four frames uniformly selected from each arterial, venous, and late phase.

Ethics approval and informed consent. This study was conducted in accordance with the Declaration of Helsinki. The use of anonymized retrospective imaging data was approved by the Institutional Review Board of The Hong Kong Polytechnic University (HSEARS20240202004). The requirement for informed consent was waived because all data were deidentified and the study involved no direct patient contact or intervention. For publicly available datasets, we complied with the data-use terms and ethics approvals of the original studies.

Data for downstream tasks. For downstream tasks related to ophthalmic disease diagnosis, we utilized seven publicly datasets. Specifically, Kaggle APTOS2019 (India), IDRID (India)[24], and MESSIDOR2 (France) were employed for diabetic retinopathy (DR) diagnosis, graded according to the International Clinical Diabetic Retinopathy Severity Scale encompassing five stages from no DR to proliferative DR. PAPILA (Spain)[25] and Glaucoma Fundus (South Korea)[31] datasets focused on glaucoma classification, spanning non-glaucoma, early glaucoma (suspected), and advanced glaucoma stages. For multi-class eye diseases, we utilized the JSIEC (China)[26] and Retina datasets. JSIEC comprises 1,000 images covering 39 fundus diseases, while the Retina dataset includes labels for normal, glaucoma, cataract, and retinal diseases. Supplementary Table 1 presents the characteristics of the datasets.

For systemic disease prediction, we used the UK Biobank dataset[27] focusing on predicting cardiovascular and neurological diseases over a 5-year period. This dataset includes 502,665 UK residents aged 40–69 years registered with the National Health Service, among whom 82,885 underwent comprehensive fundus photography (3D OCT-1000 Mark II, Topcon, Japan), resulting in 171,500 CF images. Systemic diseases evaluated include stroke, dementia, Parkinson's disease (PD), and myocardial infarction (MI). For each disease, we performed balanced sampling between the disease and healthy groups across different age ranges. We used only the CF images of

the right eye captured during a single visit per patient as input for downstream tasks to ensure consistency and minimize biases.

For the downstream retinal vessel segmentation, we used the RITE[28], which consists of 40 retinal fundus images with corresponding vessel labels. This dataset is widely used for evaluating retinal vessel segmentation algorithms.

For multi-modal image retrieval, we utilized two datasets: AngioReport[32] and Retina Image Bank[33]. The AngioReport dataset contains over 50,000 FFA and ICGA images collected from routine clinics in Thailand. We randomly selected a subset of patients from this dataset, including 387 images for our experiments. The Retina Image Bank is a large public repository from the United States, consisting of images representing 14 modalities and 84 ophthalmic diseases. We focused on images collected between 2019 and 2023, creating a custom dictionary to filter cases only involving rare diseases such as serpiginous choroiditis[39], Stargardt disease[40], retinoblastoma[41] and so on. After this selection, we identified 405 images from 150 patients for our image retrieval tasks. This subset was used to assess the model's ability to retrieve multi-modal and rare-disease cases.

Model development

Overview. Fundus2Video aims to generate a realistic FFA video $\hat{Y}$ from a given CF image x. During training, the ground-truth FFA video is represented as Y. Considering the temporal nature of the FFA series, we adopt an autoregressive GAN architecture based on pix2pixHD. The model sequentially generates each frame by conditioning on the CF image and previously generated frames. To ensure temporal consistency, we employ a multi-frame input and smoothing techniques. Specifically, three consecutive frames are used in a sliding window fashion, and a triple-frame averaging operation smooths transitions between adjacent frames.

Knowledge mask-guided video generation. To address missed fine details in critical regions, a binary knowledge mask m is generated to highlight clinically relevant changes between the early and late phases of the FFA sequence. Specifically, the mask is calculated as the thresholded difference between the first and last frames of the ground-truth FFA video Y:

$$\mathrm{m} = \delta\left(\mathrm{Y}_0 - \mathrm{Y}_\mathrm{T}\right) \tag{1}$$

where δ represents a threshold function. This mask highlights vascular and lesion regions that exhibit dynamic changes, guiding the model's attention toward clinically relevant areas.

We integrate the knowledge mask into the model through three techniques:

1. Knowledge-boosted Attention: The model's attention is guided toward focal regions using an attention-loss $\mathcal{L}_{Att}$, which compares the attention map A to the knowledge mask m:

$$\mathcal{L}_{Att} = \frac{1}{n}\sum \left(A_i - m_i\right)^2 \tag{2}$$

2. Mask-enhanced Losses $\mathcal{L}_{Mask}$: We incorporate the knowledge mask into the PatchNCE loss[42] to encourage the network to prioritize informative regions during training.
3. Knowledge-aware Discriminators: Multiple discriminators assess the images at different scales, using the mask to focus on relevant regions.

The final objective function $\mathscr{L}$ combines these techniques:

$$\mathscr{L} = \lambda_{Mask}\mathscr{L}_{Mask} + \lambda_{Att}\mathscr{L}_{Att} + \lambda_{GAN}\mathscr{L}_{GAN} \tag{3}$$

Where $\mathscr{L}_{GAN}$ is the discriminator loss, and λ_{Mask}, λ_{Att} and λ_{GAN} represent the weights assigned to different components of the loss function. These weights are set to 1, 4, and 2 to balance feature extraction, attention, and adversarial learning in the training process. An ablation study evaluating the contribution of each component is provided in Supplementary Table 8, which demonstrates that all three techniques consistently improve video generation quality, with the full model achieving the best overall performance.

Evaluation criteria

Objective evaluation criteria of video quality. Our objective evaluation criteria include Fréchet Video Distance (FVD)[43], Structural Similarity Index (SSIM)[44], Peak Signal-to-Noise Ratio (PSNR)[45], and Learned Perceptual Image Patch Similarity (LPIPS)[46]. FVD measures the similarity of feature distributions between real and generated videos, capturing overall quality and coherence. SSIM evaluates the structural similarity between the generated and ground truth videos, focusing on luminance, contrast, and structural information. PSNR quantifies reconstruction quality by comparing pixel-wise differences, with higher values indicating a closer resemblance to the ground truth. LPIPS assesses perceptual similarity using deep network features to compare image patches, capturing perceptual differences not evident through traditional pixel-based metrics. These metrics collectively ensure a thorough evaluation of the generated FFA videos, covering statistical similarity, structural fidelity, and perceptual quality. In addition to these quality-oriented metrics, we further report training time, inference time per video, and approximate TFLOPs to characterize the computational efficiency of each method under a unified setting (512 × 512 resolution, 12-frame videos). These efficiency-related metrics are particularly important for assessing the practicality of video generation models in clinical scenarios, where rapid inference and manageable computational cost are critical.

Subjective evaluation criteria of video quality. The generated videos were evaluated against real FFA videos on the test set. Two ophthalmic specialists, trained to follow a predefined consensus, assessed the videos by rating the generated videos compared to the real ones in terms of lesion integrity, retinal and choroidal structures, and dynamic range. Detailed evaluation criteria are provided in Table 2.

Downstream classification evaluation. In evaluating classification downstream tasks, we employed AUROC and AUPR metrics to assess classification performance using receiver operating characteristics and precision-recall curves, respectively. For binary classification tasks, such as the diagnosis of systemic diseases, AUROC and AUPR were computed in a binary context. For multi-class classification tasks, such as five-stage diabetic retinopathy (DR) and multi-class disease diagnoses, AUROC and AUPR were calculated individually for each disease class and then averaged (macro) to derive overall AUROC and AUPR scores.

Downstream segmentation evaluation. We evaluated vessel segmentation performance using the AUROC and Dice coefficient. AUROC measures the model's ability to distinguish between true and false positives across different thresholds, while Dice quantifies the overlap between predicted and ground truth segmentations. Higher values for both indicate better segmentation accuracy.

Downstream retrieval evaluation. For multi-modal image retrieval, we used Recall@K, where K is 1, 5, and 10, to assess the proportion of correct images retrieved within the top K results. The average recall across these values provides an overall measure of retrieval performance.

Robustness and statistical evaluation. To ensure robustness, each task underwent evaluation across 5 random splits, with models trained using different random seeds to mitigate variability. Mean performance across these iterations was computed, and 95% confidence intervals (CI) were derived using 1.96 times the standard error. We determined the statistical significance between the performance of Fundus2Video and the best-performing model among the others using two-sided t-tests.

Comparison models

Pix2pixHD[29]. Pix2pixHD is a GAN-based model designed for high-resolution image generation tasks, particularly for pixel-to-pixel translation between different modalities. It utilizes a multi-scale generator and discriminator to generate high-quality, realistic images that capture fine details. This model has been widely adopted in various fields, including

medical imaging, because it can generate clear and consistent outputs in cross-modality translation tasks.

Fundus2FA[2]. Fundus2FA is a specialized model built on pix2pixHD, designed explicitly for generating FFA images from Color Fundus CF. It is optimized for ophthalmology tasks and has been validated on downstream classification tasks related to diabetic retinopathy (DR). The model successfully generates realistic FFA images that capture essential vascular features.

RETFound[13]. RETFound is the first foundation model developed for ophthalmology, trained with separate weights for CF and Optical Coherence Tomography (OCT). Using 1.6 million unlabeled retinal images for self-supervised learning, RETFound has established itself as a versatile model capable of generalizing across various classification tasks, including disease diagnosis for retinal diseases like DR, age-related macular degeneration (AMD), glaucoma, and systemic diseases.

Medical-SAM2[30]. Medical-SAM2 is a state-of-the-art segmentation model designed to handle various medical image segmentation tasks. Medical-SAM2 has been adapted to retinal imaging to precisely segment structures like the optic disc. It employs a segmentation transformer-based architecture that excels at learning spatial dependencies and has demonstrated high performance in various medical image segmentation benchmarks. The model is recognized for its generalization capabilities, particularly in few-shot learning scenarios, making it suitable for tasks with limited labeled data.

Adaptation to downstream tasks

To evaluate the generalization capabilities of Fundus2Video, we employed its generator encoder as a feature extractor for zero-shot and few-shot classification tasks. A Multi-Layer Perceptron (MLP) was integrated with the feature extractor to perform classification. This methodology facilitates the assessment of the model's capacity to capture generalizable retinal features without extensive fine-tuning. Through comparative analysis with analogous adaptations of Fundus2FA and RETFound, we quantitatively evaluated the efficacy of our cross-modal video generation approach in learning transferable features. The MLP outputs a probability distribution over the disease categories, with the final layer containing neurons corresponding to the number of classes. In supervised training classification tasks, our primary goal is to evaluate whether Fundus2Video can enhance the performance of RETFound. To achieve this, we combined the features from the Fundus2Video generator encoder and the RETFound encoder using soft attention. The integrated features are then fed into an MLP for classification.

As for the parameter settings, for zero-shot classification, pretrained encoders were frozen and applied without task-specific fine-tuning; classification was performed using the same evaluation head across models. In the few-shot setup, we trained the models with support sets containing 1, 2, 4, or 8 examples per class. For fully supervised learning, we split the data into training, validation, and test sets with a 55:15:30 ratio. We applied label smoothing to better align the output distribution with the true labels. The training was conducted with a batch size of 16 over 50 epochs. The learning rate was gradually increased from 0 to 5×10^{-4} during the first ten epochs, followed by a cosine annealing schedule that reduced the learning rate from 5×10^{-4} to 1×10^{-6} over the remaining 40 epochs. After each epoch, model performance was evaluated on the validation set, and the model weights with the highest AUROC were saved. These best-performing weights were then used for the final test evaluations.

In the downstream task of vessel segmentation, we fine-tuned the pretrained encoders from Fundus2Video, Fundus2FA, and Pix2pixHD. We applied basic transformations such as random brightness and contrast adjustments, cropping, and flipping for augmentation. For the Medical-SAM2 model, we used the pre-trained weights from sam2_hiera_small.pt and followed the official fine-tuning protocol. In all models, the few-shot experiments were run for 100 epochs with consistent image resizing to 512 × 512 pixels. This ensured a fair comparison across methods.

For retrieval tasks, we used the same approach as zero-shot classification to extract image features. We then retrieved the top K images from the aligned latent space that were closest to a given query image. If the query result and the input image originated from the same patient, we considered the retrieval correct. We selected K values from 1, 5, 10, and reported the average recall by taking the mean of the three Recall@K scores.

Data availability

The authors do not have the right to redistribute the dataset used for training Fundus2Video due to privacy and institutional restrictions. The dataset may be available under constrained access from the corresponding author upon reasonable request. All downstream datasets used in this study are publicly accessible, including RITE (https://eye.medicine.uiowa.edu/rite-dataset), APTOS2019 (https://www.kaggle.com/c/aptos2019-blindness-detection/overview), MESSIDOR2 (https://www.adcis.net/en/third-party/messidor2/), IDRID (https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid), Glaucoma Fundus (https://doi.org/10.7910/DVN/1YRRAC), PAPILA (https://figshare.com/articles/dataset/PAPILA/14798004/1), JSIEC (https://zenodo.org/record/3477553), Retina (https://www.kaggle.com/datasets/jr2ngb/cataractdataset), UK Biobank (https://www.ukbiobank.ac.uk/), AngioReport (https://tianchi.aliyun.com/dataset/170128), and Retina Image Bank (https://imagebank.asrs.org/).

Code availability

All the code is accessible on the following GitHub repository: https://github.com/Michi-3000/Fundus2Video.

Received: 2 December 2024; Accepted: 18 May 2026;
Published online: 06 June 2026

References

1. Kornblau, I. S. & El-Annan, J. F. Adverse reactions to fluorescein angiography: A comprehensive review of the literature. *Surv. Ophthalmol.* **64**, 679–693 (2019).
2. 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).
3. 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).
4. Chen, R. et al. Noninvasive synthesis of multi-frame ultrawide-field fluorescein angiography from color fundus photographs. *Ophthalmol. Retina* 10, 176–184 (2026).
5. 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. Pract. Res.* **3**, 192–198 (2023).
6. You, A., Kim, J. K., Ryu, I. H. & Yoo, T. K. Application of generative adversarial networks (GAN) for ophthalmology image domains: a survey. *Eye Vis.* **9**, 6 (2022).
7. Ghodrati, V. et al. Temporally aware volumetric generative adversarial network-based MR image reconstruction with simultaneous respiratory motion compensation: Initial feasibility in 3D dynamic cine cardiac MRI. *Magn. Reson. Med.* **86**, 2666–2683 (2021).
8. Vukadinovic, M., Kwan, A. C., Li, D. & Ouyang, D. GANcMRI: Cardiac magnetic resonance video generation and physiologic guidance using latent space prompting. In *Machine Learning for Health (ML4H)* 594–606 (PMLR, 2023).
9. Wu, X. et al. Generation of fundus fluorescein angiography videos for health care data sharing. *JAMA Ophthalmol.* **143**, 623–632 (2025).
10. Yu, T. et al. A systematic review of advances in AI-assisted analysis of fundus fluorescein angiography (FFA) images: from detection to report generation. *Ophthalmol. Ther.* **14**, 599–619 (2025).

11. Shao, A. et al. Generative artificial intelligence for fundus fluorescein angiography interpretation and human expert evaluation. *npj Digit. Med.* **8**, 396 (2025).
12. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
13. Zhou, Y. et al. A foundation model for generalizable disease detection from retinal images. *Nature* **622**, 156–163 (2023).
14. Shi, D. et al. EyeFound: a multimodal generalist foundation model for ophthalmic imaging. *arXiv preprint arXiv:2405.11338* (2024).
15. Qiu, J. et al. Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence. *NEJM AI* **1**, AIoa2300221 (2024).
16. Shi, D. et al. A multimodal visual–language foundation model for computational ophthalmology. *npj Digit. Med.* **8**, 381 (2025).
17. Silva-Rodriguez, J., Chakor, H., Kobbi, R., Dolz, J. & Ayed, I. B. A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. *Med. Image Anal.* **99**, 103357 (2025).
18. Cheng, D. et al. SAM on medical images: a comprehensive study on three prompt modes. *arXiv preprint arXiv:2305.00035* (2023).
19. Morano, J. et al. Multimodal foundation model and benchmark for comprehensive retinal OCT image analysis. *NPJ Digit. Med.* **8**, 576 (2025).
20. Pan, J. et al. Video generation from single semantic label map. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 3733–3742 (2019).
21. Dorjsembe, Z., Pao, H.-K., Odonchimed, S. & Xiao, F. Conditional diffusion models for semantic 3D brain MRI synthesis. *IEEE J. Biomed. Health Inform*. **28**, 4084–4093 (2024).
22. Ren, W. et al. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324* (2024).
23. Henschel, R. et al. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In *Proc. Computer Vision and Pattern Recognition Conference* 2568–2577 (2025).
24. Porwal, P. et al. Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. *Data* **3**, 25 (2018).
25. 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).
26. 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).
27. Chua, S. Y. L. et al. Cohort profile: design and methods in the eye and vision consortium of UK Biobank. *BMJ Open* **9**, e025077 (2019).
28. Hu, Q., Abramoff, M. D. & Garvin, M. K. Automated separation of binary overlapping trees in low-contrast color retinal images. *Med. Image Comput Comput. Assist. Inter.* **16**, 436–443 (2013).
29. Wang, T.-C. et al. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 8798–8807 (2018).
30. Zhu, J., Hamdi, A., Qi, Y., Jin, Y. & Wu, J. Medical Sam 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874* (2024).
31. 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).
32. Xu, P. et al. AngioReport: dataset and baseline methods for fundus angiography report generation. *Br. J. Ophthalmol.* **109**, 1283–1288 (2025).
33. Retina Image Bank, available at https://imagebank.asrs.org/.
34. Sabanayagam, C. et al. Incidence and progression of diabetic retinopathy: a systematic review. *Lancet Diab Endocrinol.* **7**, 140–149 (2019).
35. Stein, J. D., Khawaja, A. P. & Weizer, J. S. Glaucoma in adults-screening, diagnosis, and management: a review. *JAMA* **325**, 164–174 (2021).
36. 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).
37. Wu, B. et al. Visual transformers: Token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677* (2020).
38. Liu, X. et al. Toward the unification of generative and discriminative visual foundation model: a survey. *Visual Comput.* **41**, 3371–3412 (2024).
39. Lee, S. S. & Mackey, D. A. Glaucoma - risk factors and current challenges in the diagnosis of a leading cause of visual impairment. *Maturitas* **163**, 15–22 (2022).
40. Tanna, P., Strauss, R. W., Fujinami, K. & Michaelides, M. Stargardt disease: clinical features, molecular genetics, animal models and therapeutic options. *Br. J. Ophthalmol.* **101**, 25–30 (2017).
41. Dimaras, H. et al. Retinoblastoma. *Lancet* **379**, 1436–1446 (2012).
42. Li, F., Hu, Z., Chen, W. & Kak, A. Adaptive supervised patchnce loss for learning h&e-to-ihc stain translation with inconsistent groundtruth image pairs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 632–641 (Springer, 2023).
43. Unterthiner, T. et al. FVD: A new metric for video generation. In *DeepGenStruct Workshop at ICLR* (2019).
44. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004).
45. Huynh-Thu, Q. & Ghanbari, M. Scope of validity of PSNR in image/video quality assessment. *Electron. Lett.* **44**, 800–801 (2008).
46. Zhang, R., Isola, P., Efros, A. A., Shechtman, E. & Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 586–595 (2018).
47. Zhang, W. et al. Fundus2Video: Cross-Modal Angiography Video Generation from Static Fundus Photography with Clinical Knowledge Guidance. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 689–699 (Springer, 2024).

Acknowledgements

This study was supported by the Start-up Fund for RAPs under the Strategic Hiring Scheme (P0048623) from HKSAR, the Global STEM Professorship Scheme (P0046113), and the 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. We also thank the InnoHK HKSAR Government for providing valuable support. The research work described in this paper was conducted in the JC STEM Lab of Innovative Light Therapy for Eye Diseases funded by The Hong Kong Jockey Club Charities Trust. This study is an extension of our previous publication accepted to the MICCAI 2024 conference[47].

Author contributions

W.Z., J.Y., S.H., D.S., and M.H. conceived the study. W.Z. and D.S. developed the model and performed the experiments. M.H. provided data and computing resources. W.Z., J.Y., R.C., S.H., P.X., Z.W., and D.S. contributed to data interpretation. W.Z. drafted the manuscript. W.Z., J.Y., R.C., S.H., P.X., X.C., Z.W., S.L., H.C., X.S., D.S., and M.H. critically reviewed and revised the manuscript. D.S. and M.H. supervised the project. All authors read and approved the final manuscript.

Competing interests

A patent application has been filed pertaining to the results presented in this paper. The patent (Application No.: CN 202410360491.4) is pending with The Hong Kong Polytechnic University as the applicant, and Mingguang He, Danli Shi as inventors. The patent relates to the cross-modality generative method and its applications described in this manuscript. Jiancheng Yang is

an Editorial Board Member of *npj Digital Medicine*. He was not involved in the peer review or decision-making process for this manuscript. The remaining authors declare no competing financial or non-financial interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-026-02804-1.

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

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

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

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

© The Author(s) 2026