FFA Sora: generating fundus fluorescein angiography videos for healthcare data sharing

Authors:

Xinyuan Wu, MD[1#], Lili Wang, MS[2#], Ruoyu Chen, MD[1#], Bowen Liu, MS[1], Weiyi Zhang, MS[1], Xi Yang, MD, PhD[3], Yifan Feng, MD, PhD[3], Mingguang He, MD, PhD[1,4,5] and Danli Shi, MD, PhD[1,4]✉

Affiliations:

1. School of Optometry, The Hong Kong Polytechnic University, Hong Kong.

2. Department of Computing, The Hong Kong Polytechnic University, Hong Kong.

3. Department of Ophthalmology, Zhongshan Hospital, Fudan University, Shanghai, China.

4. Research Centre for SHARP Vision (RCSV), The Hong Kong Polytechnic University, Hong Kong.

5. Centre for Eye and Vision Research (CEVR), 17W Hong Kong Science Park, Hong Kong.

# Xinyuan Wu, Lili Wang and Ruoyu Chen contributed equally to this work.

Correspondence:

Dr. Danli Shi, The Hong Kong Polytechnic University, Hong Kong, China. Email: danli.shi@polyu.edu.hk

Prof. Mingguang He, Chair Professor of Experimental Ophthalmology, The Hong Kong

Polytechnic University, Hong Kong, China. Email: mingguang.he@polyu.edu.hk

Manuscript word count: 2995 words**.**

Running title: FFA Sora generates fundus fluorescein angiography videos

Key Points:

Question Can artificial intelligence (AI) models generate dynamic fundus fluorescein angiography (FFA) videos directly from textual descriptions of retinal conditions?

Findings This study developed FFA Sora, an AI-driven text-to-video model capable of generating high-quality FFA videos from medical textual descriptions. The generated videos depicted retinal abnormalities, such as non-perfusion areas, leakage, and microaneurysms. Objective and subjective assessments confirmed their alignment with corresponding lesion or diseases descriptions. Retrieval Evaluation demonstrated the model's privacy-preserving performance.

Meaning AI-driven text-to-video models offer an approach to visualizing various retinal conditions from text prompts, potentially promoting privacy-preserving data sharing and improving comprehension in medical education.

Abstract

Importance Medical data sharing faces strict restrictions. Text-to-video generation shows potential for creating realistic medical data while preserving privacy, offering a solution for cross-centre data sharing and medical education.

Objective To develop and evaluate a text-to-video generative model that converts report texts into dynamic FFA videos, enabling visualization of retinal vascular and structural abnormalities.

Design, Setting, and Participants This study retrospectively collected anonymized FFA data from a tertiary hospital in China, including 3625 FFA videos. Using this dataset, a text-to-video model named FFA Sora was developed and evaluated, with 80% for training, 10% for validation and the remaining 10% for testing. The model integrates two key components: the Wavelet-Flow Variational Autoencoder (WF-VAE) and the diffusion transformer (DiT).

Main Outcomes and Measures The model's performance was assessed through objective metrics, including Fréchet Video Distance (FVD), Learned Perceptual Image Patch Similarity (LPIPS), and Visual-question-answering Score (VQAScore). FFA domain-specific evaluation was measured by Bidirectional Encoder Representations from Transformers Score (BERTScore), and image retrieval was evaluated by Recall@K. Human evaluations were conducted on a scale of 1 (best) to 5 (worst).

Results The generated FFA videos demonstrated retinal abnormalities from the input text, as

confirmed by objective metrics: FVD = 2273, LPIPS = 0.48±0.043, and VQAScore = 0.61 ±0.078. Domain-specific evaluations showed alignment between the generated videos and textual prompts, with BERTScore of 0.35±0.091. Additionally, the model demonstrated privacy-preserving performance in retrieval evaluations, achieving an average Recall@K of 0.073. Human assessments indicated visual quality, with an average score of 1.570±0.440.

Conclusions and Relevance This study demonstrated that FFA Sora, an AI-driven text-to-video model, generated FFA videos from textual descriptions, potentially improving visualization for clinical and educational purposes. Its privacy-preserving design may address key challenges in data sharing while trying to ensure compliance with confidentiality standards.

Introduction

Fundus fluorescein angiography (FFA) enables the dynamic visualization of retinal blood flow and lesional changes through the intravenous injection of fluorescein sodium. This technique can be important for assessing blood-retina barrier function and diagnosing various retinal vascular diseases, including diabetic retinopathy (DR) and choroidal neovascularization (CNV).[1-5] Despite its importance, accurately interpreting FFA images demands expertise. For ophthalmologists and medical students, the challenge of correlating textual description on text book with corresponding FFA findings may be compounded by the complexity of disease presentations and the limited availability of representative FFA cases.[6]

The growing application of artificial intelligence (AI) offers a potential solution for automatic image interpretation. However, developing and validating these tools requires extensive, high-quality labeled datasets, which can be difficult to obtain. Strict regulations governing the use and exchange of personal health information protect patient privacy but can inadvertently hinder the collaborative data sharing necessary to build effective AI models.[7] Additionally, healthcare providers often work in isolation, making them reluctant to share the clinical data for boosting AI model training. Therefore, innovative strategies are needed to overcome the challenges.

Open AI's Sora has been reported to generate realistic videos based on text inputs,[8] paving the way for the synthesis of well-labeled, diversified, and privacy-preserving medical data. Recently, diffusion models (DMs) have gained attention for their capacity to generate high-

quality, temporally consistent images and videos through iterative denoising processes.[9] These models have been applied to generate medical data, including chest X-ray images[10], brain MRI scans,[11] multi-modal retinal images,[12] endoscopy videos,[13] and facilitate neurosurgery.[14] These technological advancements have expanded applications across multiple domains, including medical education, clinical training, surgical simulation, medical knowledge dissemination, doctor-patient communication, and biomedical data augmentation.[8] However, generating videos from textual description has yet to be explored for creating lesion-preserving dynamic FFA videos in ophthalmology.

To bridge the gap, we developed FFA Sora, an AI-driven model capable of producing dynamic FFA videos from textual reports. This cross-modality generative model might assist in privacy-preserving data sharing and medical education.

Methods

We utilized de-identified existing data for our study, which received approval from the Institutional Review Board of the Hong Kong Polytechnic University. Our study adhered to Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines.

Dataset

This study uses a subset of a retrospective FFA dataset.[15] FFA data were obtained from a tertiary hospital in China. The dataset included both the medical records and FFA examinations of

patients assessed between November 2016 and December 2019. All FFA videos were captured using Zeiss FF450 Plus and Heidelberg Spectralis systems (Heidelberg, Germany) at a resolution of 768×768 pixels. The collection encompassed a wide variety of ocular conditions, such as DR, retinal vein occlusion (RVO), and central serous chorioretinopathy (CSC). To try to ensure high-quality data, we implemented a filtering process based on vessel area ratios within the FFA frames, excluding any video whose frame vessel area ratios fell below 0.005.

To address the issue of inconsistent and insufficient frame counts in the original FFA videos, we standardized all videos to a fixed frame count of 21 frames. For videos with more than 21 frames, we selected 21 frames in reverse chronological order. For videos with fewer than 21 frames, additional frames were interpolated to reach the required count. This was achieved using a frame interpolation technique based on linear interpolation, implemented via the OpenCV library.

The process involved extracting all frames from the input video and computing the interpolation positions based on the target frame count. For each interpolated frame, a weighted linear interpolation between two adjacent frames was employed to generate smooth transition frames. Specifically, let an interpolated frame lie between the i -th frame and the (i+1)-th frame of the original video. The pixel values of the interpolated frame are computed as follows:

$$F_{interpolated} = (1 - t) \times F_i - t \times F_{i+1}$$

where:

F_i and F_{i+1} denote the pixel values of the i -th and the (i+1)-th frames, respectively, t is the interpolation ratio, which takes values in the range [0,1] and determines the relative contribution of the two adjacent frames to the interpolated frame.

The interpolation ratio t is calculated based on the relative temporal position of the interpolated frame within the sequence. Once all interpolated frames were generated, they were combined with the original frames to produce a standardized video consisting of exactly 21 frames. All patient data were stratified by patient ID and randomly divided into 80% for training, 10% for validation and 10% for testing.

Model Architecture

The framework of FFA Sora was based on the Open-Sora Plan repository (v1.3.0),[16] which we specifically adapted to address the task of generating FFA images from medical textual descriptions. It consisted of two primary components: the Wavelet-Flow Variational Autoencoder (WF-VAE) and the Diffusion Transformer (DiT). Both were trained on a single NVIDIA A800 80,000 MB (80 GB) PCIe GPU with CUDA 12.1. The programming environment used for this project is Python 3.8.20 with PyTorch 2.1.0 (CUDA 12.1).

The WF-VAE architecture[17], derived from the Stable-Diffusion Image VAE, was fine-tuned and optimized for the specialized task of text-to-FFA image generation in the medical domain. A key architectural enhancement involved the transformation of Conv2D layers into CausalConv3D layers, enabling the model to process temporal video data, which was essential

for the sequential nature of FFA imaging. Furthermore, multi-level wavelet transforms were employed in the encoder to decompose video signals into frequency sub-bands, allowing low-frequency components, which is critical for FFA imaging details, to bypass the backbone network via the Main Energy Flow Pathway, thereby reducing computational redundancy. The model was trained with a latent dimension of 8 for 200,000 steps, with GPU memory consumption reaching approximately 19,954 MB.

The second major component was the DiT architecture,[18] which employed a Transformer-based architecture with cross-modal attention mechanisms to optimize text-to-FFA image generation. The denoising process was performed in a low-dimensional latent space compressed by the VAE, where each Cross-DiT Block, consisting of self-attention, cross-attention, and feed-forward networks, was enhanced with gating mechanisms to refine feature extraction and fusion. This design improved the model's ability to capture both high-level semantics and fine-grained details, ensuring that the generated FFA images accurately reflected clinically relevant features such as global vascular structures and local pathological regions. The fine-tuning process was conducted on the same dataset for 50,000 steps over approximately 72 hours, with GPU memory consumption reaching approximately 73,728 MB.

Model Inference

During the inference phase, the model processed input textual descriptions and generated corresponding FFA images with a resolution of 21×512×512. The input text typically contained detailed medical descriptions of retinal conditions, such as microaneurysms, non-perfusion area,

leakage, and other abnormalities, which were crucial for FFA-based diagnosis. To ensure that the generated images align closely with the described retinal conditions, our model utilized the pre-trained text encoder to capture the semantic nuances of the input effectively. These latent features, conditioned on the text, were passed through the CausalConv3D layers in the decoder, which reconstructed the FFA images by generating temporal and spatial patterns consistent with real FFA images. The inference process employed a tiled convolution approach, which optimized memory usage and ensured high-quality image output. By applying this technique, the model generated coherent and clinically relevant FFA images that accurately depict the specified retinal conditions, while maintaining computational efficiency during inference. The overview of FFA Sora is presented in Figure 1.

Objective Evaluation

Several objective metrics were employed to assess the quality of the generated FFA videos, including Fréchet Video Distance (FVD),[19] Learned Perceptual Image Patch Similarity (LPIPS),[20] and Visual-question-answering Score (VQAScore).[21] FVD discerns the likeness in feature distribution between authentic and synthetic videos, encapsulating the holistic excellence and unity. FVD typically ranges from 0 to higher values, with lower values indicating better similarity. LPIPS, utilizing deep learning-inspired features, measures the perceptual resemblance by comparing image segments, thus potentially revealing nuances beyond the capabilities of conventional pixel-based assessments. LPIPS scores range from 0 to 1, with lower scores considered higher perceptual similarity. VQAScore evaluates the alignment of text-to-video models. This score ranges from 0 to 1, where higher values represent

better quality. Collectively, these evaluation criteria potentially provide a comprehensive analysis of the generated FFA video quality.

For FFA domain-specific evaluation, we used FFA-GPT to translate the generated FFA video into text report and compared it with the original text prompt using Bidirectional Encoder Representations from Transformers Score (BERTScore)[22] as similarity measure. BERTScore utilizes the pre-trained BERT language model to measure conceptual overlap between generated and reference content. Unlike purely lexical metrics, it ideally captures nuanced alignments within broader contexts, providing a deeper assessment of textual congruence. This score ranges from 0 to 1, with higher values indicating better alignment between the generated and original content.

For privacy preservation evaluation, we assessed FFA Sora's ability to generate content-preserving yet deidentifiable data. We employed an image-to-image retrieval method, using Recall@K[23] to measure the proportion of matched visuals retrieved among the top K results. It ranges from 0 to 1, with lower values indicating that the original lesion information cannot be captured in the generated videos. Since the accuracy of the generated videos can be validated through other parameters mentioned above, Recall@K serves as an indicator of FFA Sora's ability in preserving privacy.

Human Assessment

This evaluation followed the methodology described previously.[24,25] Three ophthalmologists

(X.W., X.Y. and Y.F.) with 3, 5, and 10 years of experience respectively, representing trainee, attending, and expert levels, reviewed 50 randomly selected FFA videos generated from the test set, comparing them against the corresponding ground-truth FFA videos. Each video was rated on a scale from 1 to 5, where 1 represented excellent quality and 5 indicated very poor quality. The specific criteria were defined in eFigure 3 in the Supplement.

Results

The final dataset used in the study consists of 3625 FFA videos, among which 2851 videos were randomly selected for training, 387 videos for validation, and 387 videos for testing. Demographic characteristics of the dataset were presented in eTable1 in the Supplement. FFA reports in the dataset included description of various lesions, such as microaneurysms, leakage, neovascularization, capillary non-perfusion, and macular edema (eTable 2 in the Supplement).

Figure 2 demonstrates representative examples of generated FFA videos, including corresponding prompts, ground-truth videos, and the generated video frames. FFA Sora could generate detailed FFA videos from text prompts, illustrating retinal lesions and vascular abnormalities, such as leakage, neovascularization, and microaneurysms, etc. Our model could produce FFA videos for other retinal and choroidal diseases as well, such as uveitis, retinitis pigmentosa (RP), etc. Representative frames of these videos are shown in eFigure 1 and eFigure 2 in the Supplement.

Objective evaluation showed excellent quality of generated FFA videos

The generated videos were comprised of 21 frames that precisely document the entire process of FFA examination, with a particular emphasis on the venous and late phases. The model we employed exhibited satisfactory performance in producing FFA videos when applied to the test datase, as evidenced by FVD = 2273, LPIPS = 0.48±0.043, and VQAScore = 0.61±0.078 (Figure 4A). These metrics served as benchmarks in assessing the fidelity and quality of the generated FFA videos.

In addition to the evaluation with above standard generative metrics, we also proposed domain-specific evaluation strategies. Based on previous development of FFA-GPT,[15] a two-stage system which can generate FFA report automatically, our model was evaluated to determine whether the content of the generated videos aligns with the actual characteristics of retinal diseases as denoted by specified textual prompts, with BERTScore = 0.35±0.091 (Figure 4A). To visualize these findings, an example of reports by FFA-GPT based on FFA Sora's generated FFA videos was presented in Figure 3. Additionally, Figure 4C demonstrated an example where the objective evaluation results were relatively poor, with ranking in the lowest 20% across all parameters.

Image retrieval demonstrated FFA Sora's excellent performance on privacy preserving

To further explore FFA Sora's robust feature representation, we investigated its performance in image retrieval tasks, specifically examining the relationship between the generated videos and their ground-truth counterparts. The results of this investigation were summarized in Figure 4B, where we reported Recall@K scores that include 0.02, 0.04, 0.16 for K = 5, 10, and 50,

respectively. Furthermore, the mean recall was calculated to be 0.073, reflecting disparities among generated and ground-truth videos. These relatively low Recall@K values indicated that FFA Sora could prevent the leakage of confidential image information while reserving the essential pathological features during video generation.

Human assessment confirmed high visual quality of generated FFA videos

A comprehensive analysis was conducted on fifty randomly selected videos produced via our model, which were subjected to meticulous review by three ophthalmologists. The generated videos were systematically compared with the corresponding textual reports and the ground-truth FFA videos to rigorously assess their accuracy and fidelity. The subjective visual quality score averaged at 1.570±0.440, with a Kappa value of 0.761. The specific results are presented in Figure 5A and 5B, while Figure 5C illustrated an example where the generated video received consistently low subjective ratings. To clarify the scoring scale, examples corresponding to ratings from 1 to 5 were presented in eFigure 3 in the Supplement. Additionally, eFigure 4 in the Supplement demonstrated cases where discrepancies arose between objective and subjective evaluations.

Discussion

In this study, we developed FFA Sora, a text-to-video diffusion transformer designed to generate FFA videos from textual descriptions. This research demonstrated the potential of text-to-video models for visualizing FFA-related terms and associated manifestations of several retinal disease, offering a solution for privacy-preserving data sharing in machine learning model

297 training and medical AI applications.

298

299 Generative models represent potential advancements in medical data generation of their ability
300 to produce images of high quality, stability, and diversity.[26-28] Previous research has explored
301 applications of generative adversarial networks (GANs) in retinal image quality enhancement,
302 domain adaptation, vessel segmentation, and cross-modality generation of fundus
303 autofluorescence, FFA, and indocyanine green angioagraphy.[29-35] More recently, diffusion
304 models beat GANs in image generation[9] and have been applied to ocular imaging tasks, such
305 as ocular surface structure segmentation for meibomian gland dysfunction grading[36] and data
306 augmentation using retinography images to support deep learning model training.[37] DiT, a
307 transformer-based architecture for diffusion models, enhances scalability and generative
308 performance by replacing the conventional U-Net backbone with a transformer. Its architectural
309 innovation improves feature representation, synthesis quality, and cross-modality
310 adaptablitily.[38-40] Building on this foundation, FFA Sora is an advanced application of
311 generative models with a potentially better architectural extensibility. Compared with our
312 previous image-to-video study[24,25], FFA Sora exhibits higher LPIPS and FVD values because
313 text-based video generation introduces variations in lesion presentation, whereas image-based
314 video generation retains the overall structural integrity of the original input. Comprehensive
315 evaluations confirmed that the produced videos preserved diagnostic plausibility when visually
316 diversed.

317

318 Data sharing in AI model training is often hindered by rigorous privacy regulations and

concerns regarding patient confidentiality, particularly with sensitive medical imaging and clinical videos.[41,42] These challenges limit collaboration and the creation of large, diverse datasets needed for training and validating AI models. Additionally, there is a risk that AI models may inadvertently generate videos by recalling data from their training set, potentially compromising privacy.[43] The notably low Recall@K scores in our image retrieval evaluation demonstrated the visual differences between synthetic videos and the ground-truth ones, thus preserving privacy disease information. The visual distinction in generated outputs ensures that the resulting datasets are not mere replications of existing clinical images, allowing for simulating the diversity in real-world cases of the same disease. By mitigating these concerns, FFA Sora facilitates cross-center data sharing and reproducibility, paving the way for potentially improving collaborative research, clinical training, and the development of AI-driven diagnostic tools.

AI-generated content in the medical domain has broad implications. Research on AI-assisted medical image generation has demonstrated its value in medical education[6] and the training of convolutional neural networks (CNNs) for medical purposes such as diagnosis and disease classification.[10,44] For instance, Tabuchi H, et al.[6] introduced a text-to-image model to produce ultra-widefield (UWF) retinal images, significantly improving diagnostic accuracy among medical students. This outcome highlights the potential of integrating AI technologies into medical education. Compared with their model, FFA Sora offers a more dynamic demonstration of retinal anatomy, vascular perfusion, and pathology, thereby facilitating ophthalmology training. FFA Sora can simulate rare and complex pathologies, offering an interactive learning

tool to improve diagnostic skills, decision-making, and treatment planning, particularly in resource-limited settings. FFA Sora can also contribute to case libraries for board exams and clinical assessments, providing trainees with diverse retinal conditions that may not be encountered during rotations. Its accessibility potentially ensures its broad applicability across educational institutions and healthcare facilities, potentially equpping future ophthalmologists with AI-driven tools for diagnosing and managing retinal diseases. Additionally, FFA Sora has potential in clinical workflows, particularly for patient education and disease progression simulation. For example, clinicians counseling patients with severe NPDR can use AI-generated FFA videos to illustrate disease progression. This dynamic visualization can help patients understand the impact of uncontrolled diabetes and the benefits of timely interventions, potentially improving adherence to treatments such as blood sugar control, laser therapy, and anti-VEGF injections.

Limitations

This research has several limitations. First, although FFA Sora demonstrates high accuracy and reliability, its real-world clinical utility requires further validation. Broader external validation using diverse datasets would enhance its practical relevance. Second, our evaluation approach may not fully capture the generative quality. Our evaluators were not masked to the synthetic nature of the generated videos. Future work should incorporate masked evaluations to better quantify realism and ensure synthetic videos closely resemble real-world data. Finally, our current approach focuses solely on visualizing FFA reports, without integrating other imaging modalities. Multimodal imaging could improve AI-based models in ophthalmology, offering

more complete diagnostic and treatment support.

Conclusion

In conclusion, this study presents FFA Sora, a model that transforms FFA report text into dynamic video representations. These synthetic videos appear to retain critical clinical details while preserving patient privacy, offering a potential solution to data-sharing challenges in healthcare. By enabling potentially secure and efficient visualization, FFA Sora may support collaborative research, enhance medical education, and accelerate the development of AI-driven diagnostic and medical systems.

Acknowledgement

We thank the InnoHK HKSAR Government for providing valuable supports.
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.

Funding

M.H. disclose support for the research and publication of this work from the Global STEM Professorship Scheme (P0046113), and Henry G. Leong Endowed Professorship in Elderly Vision Health.

Role of the Funder/Sponser

The funders had no role in the design and conduct of the study, collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and

decision to submit the manuscript for publication.

Author contributions

D.S. conceived the study. L.W. built the text-to-video model. D.S., R.C., B.L., W.Z. conducted the literature search, analyzed the data. L.W. performed objective evaluation. X.W., X.Y., Y.F. completed visual evaluation. X.W., L.W. wrote the manuscript. L.W., X.W. organized figures and tables in this study. M.H. provided the data and facilities. All authors critically revised the manuscript.

Confict of Interest Disclosures

The authors declare no conflict interest.

Access to data and data analysis

Code is available at https://github.com/PKU-YuanGroup/Open-Sora-Plan/tree/v1.3.0.

The generated dataset will be available on request, but the original dataset involving human participants will remain confidential.

Reference

1. Burns SA, Elsner AE, Gast TJ. Imaging the Retinal Vasculature. *Annual Review of Vision Science*. 2021-09-15 2021;7:129-153. doi:10.1146/annurev-vision-093019-113719
2. Sulzbacher F, Pollreisz A, Kaider A, Kickinger S, Sacu S, Schmidt-Erfurth U. Identification and clinical role of choroidal neovascularization characteristics based on optical coherence tomography angiography. *Acta Ophthalmol*. Jun 2017;95(4):414-420. doi:10.1111/aos.13364
3. Do DV. Detection of new-onset choroidal neovascularization. *Curr Opin Ophthalmol*. May 2013;24(3):244-7. doi:10.1097/ICU.0b013e32835fd7dd
4. Coscas GJ, Lupidi M, Coscas F, Cagini C, Souied EH. OPTICAL COHERENCE TOMOGRAPHY ANGIOGRAPHY VERSUS TRADITIONAL MULTIMODAL IMAGING IN ASSESSING THE ACTIVITY OF EXUDATIVE AGE-RELATED MACULAR DEGENERATION: A New Diagnostic Challenge. *Retina*. Nov 2015;35(11):2219-28. doi:10.1097/iae.0000000000000766
5. Sabanayagam C, Banu R, Chee ML, et al. Incidence and progression of diabetic retinopathy: a systematic review. *Lancet Diabetes Endocrinol*. Feb 2019;7(2):140-149. doi:10.1016/s2213-8587(18)30128-1
6. Tabuchi H, Engelmann J, Maeda F, et al. Using artificial intelligence to improve human performance: efficient retinal disease detection training with synthetic images. *The British Journal of Ophthalmology*. 2024-09-20 2024;108(10):1430-1435. doi:10.1136/bjo-2023-324923
7. Price WN, 2nd, Cohen IG. Privacy in the age of medical big data. *Nat Med*. Jan 2019;25(1):37-43. doi:10.1038/s41591-018-0272-7
8. O'Callaghan J. How OpenAI's text-to-video tool Sora could change science – and society. *Nature*. 2024-03-12 2024;627(8004):475-476. doi:10.1038/d41586-024-00661-0
9. Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*. 2021;34:8780-8794.
10. Bluethgen C, Chambon P, Delbrouck J-B, et al. A vision-language foundation model for the generation of realistic chest X-ray images. *Nature Biomedical Engineering*. 2024-08-26 2024;doi:10.1038/s41551-024-01246-y
11. Dorjsembe Z, Pao H-K, Odonchimed S, Xiao F. Conditional Diffusion Models for Semantic 3D Brain MRI Synthesis. *IEEE Journal of Biomedical and Health Informatics*. 7/2024 2024;28(7):4084-4093. doi:10.1109/JBHI.2024.3385504
12. Chen R, Zhang W, Liu B, et al. EyeDiff: text-to-image diffusion model improves rare eye disease diagnosis. *arXiv preprint arXiv:241110004*. 2024;
13. Li C, Liu H, Liu Y, et al. Endora: Video Generation Models as Endoscopy Simulators. 2024-03-17 2024;
14. Sener F, Saraf R, Yao A. Transferring Knowledge From Text to Video: Zero-Shot Anticipation for Procedural Actions. *IEEE transactions on pattern analysis and machine intelligence*. 2023-06 2023;45(6):7836-7852. doi:10.1109/TPAMI.2022.3218596
15. Chen X, Zhang W, Xu P, et al. FFA-GPT: an automated pipeline for fundus fluorescein angiography interpretation and question-answer. *npj Digital Medicine*. 2024-05-03 2024;7(1):1-9. doi:10.1038/s41746-024-01101-z
16. Lin B, Ge Y, Cheng X, et al. Open-Sora Plan: Open-Source Large Video Generation Model. 2024:
17. Li Z, Lin B, Ye Y, et al. WF-VAE: Enhancing Video VAE by Wavelet-Driven Energy Flow for

Latent Video Diffusion Model. 2024:
18. Peebles WS, Xie S. Scalable Diffusion Models with Transformers. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2022:4172-4182.
19. Unterthiner T, van Steenkiste S, Kurach K, Marinier R, Michalski M, Gelly S. FVD: A new metric for video generation. 2019;
20. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. presented at: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2018; https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00068
21. Lin Z, Pathak D, Li B, et al. Evaluating Text-to-Visual Generation with Image-to-Text Generation. 2024:
22. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating Text Generation with BERT. *ArXiv*. 2019;abs/1904.09675
23. Song HO, Xiang Y, Jegelka S, Savarese S. Deep Metric Learning via Lifted Structured Feature Embedding. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015:4004-4012.
24. Zhang W, Yang J, Chen R, et al. Fundus to Fluorescein Angiography Video Generation as a Retinal Generative Foundation Model. *ArXiv*. 2024;abs/2410.13242
25. Zhang W, Huang S, Yang J, et al. Fundus2Video: Cross-Modal Angiography Video Generation from Static Fundus Photography with Clinical Knowledge Guidance. Springer Nature Switzerland; 2024:689-699.
26. Tian Y, Fan L, Isola P, Chang H, Krishnan D. StableRep: Synthetic Images from Text-to-Image Models Make Strong Visual Representation Learners. *ArXiv*. 2023;abs/2306.00984
27. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021:10674-10685.
28. Croitoru FA, Hondru V, Ionescu RT, Shah M. Diffusion Models in Vision: A Survey. *IEEE Trans Pattern Anal Mach Intell*. Sep 2023;45(9):10850-10869. doi:10.1109/tpami.2023.3261988
29. He S, Joseph S, Bulloch G, et al. Bridging the Camera Domain Gap With Image-to-Image Translation Improves Glaucoma Diagnosis. *Transl Vis Sci Technol*. 2023;12(12):20-20. doi:10.1167/tvst.12.12.20
30. 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*. Nov-Dec 2023;3(4):192-198. doi:10.1016/j.aopr.2023.11.001
31. Shi D, He S, Yang J, Zheng Y, He M. One-shot Retinal Artery and Vein Segmentation via Cross-modality Pretraining. *Ophthalmol Sci*. Mar-Apr 2024;4(2):100363. doi:10.1016/j.xops.2023.100363
32. Chen R, Xu K, Zheng K, et al. Generating Multi-frame Ultrawide-field Fluorescein Angiography from Ultrawide-field Color Imaging Improves Diabetic Retinopathy Stratification. *arXiv preprint arXiv:240810636*. 2024;
33. Shi D, Zhang W, He S, et al. Translation of Color Fundus Photography into Fluorescein Angiography Using Deep Learning for Enhanced Diabetic Retinopathy Screening. *Ophthalmol Sci*. Dec 2023;3(4):100401. doi:10.1016/j.xops.2023.100401
34. Zhang W, Huang S, Yang J, et al. Fundus2Video: Cross-Modal Angiography Video Generation from Static Fundus Photography with Clinical Knowledge Guidance. Springer Nature Switzerland;

2024:689-699.
35. Chen R, Zhang W, Song F, et al. Translating color fundus photography to indocyanine green angiography using deep-learning for age-related macular degeneration screening. *npj Digital Medicine*. 2024-02-12 2024;7(1):34. doi:10.1038/s41746-024-01018-7
36. Guo X, Wen H, Hao H, et al. Randomness-restricted Diffusion Model for Ocular Surface Structure Segmentation. *IEEE Trans Med Imaging*. Nov 11 2024;Ppdoi:10.1109/tmi.2024.3494762
37. Aktas B, Ates DD, Duzyel O, Gumus A. Diffusion-based data augmentation methodology for improved performance in ocular disease diagnosis using retinography images. *International Journal of Machine Learning and Cybernetics*. 2024/12/11 2024;doi:10.1007/s13042-024-02485-w
38. Peebles W, Xie S. Scalable diffusion models with transformers. 2023:4195-4205.
39. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv*. 2020;abs/2010.11929
40. Bao F, Nie S, Xue K, et al. All are Worth Words: A ViT Backbone for Diffusion Models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022:22669-22679.
41. Yang Y, Chen X, Lin H. Privacy preserving technology in ophthalmology. *Current Opinion in Ophthalmology*. 11/2024 2024;35(6):431-437. doi:10.1097/ICU.0000000000001087
42. Aa S, M G, C A-R, et al. Privacy-preserving federated machine learning on FAIR health data: A real-world application. *Computational and structural biotechnology journal*. 02/17/2024 2024;24doi:10.1016/j.csbj.2024.02.014
43. Collado-Mesa F, Alvarez E, Arheart K. The Role of Artificial Intelligence in Diagnostic Radiology: A Survey at a Single Radiology Residency Training Program. *J Am Coll Radiol*. Dec 2018;15(12):1753-1757. doi:10.1016/j.jacr.2017.12.021
44. Coyner AS, Chen JS, Chang K, et al. Synthetic Medical Images for Robust, Privacy-Preserving Training of Artificial Intelligence: Application to Retinopathy of Prematurity Diagnosis. *Ophthalmology Science*. 2022-06 2022;2(2):100126. doi:10.1016/j.xops.2022.100126

Figure legends

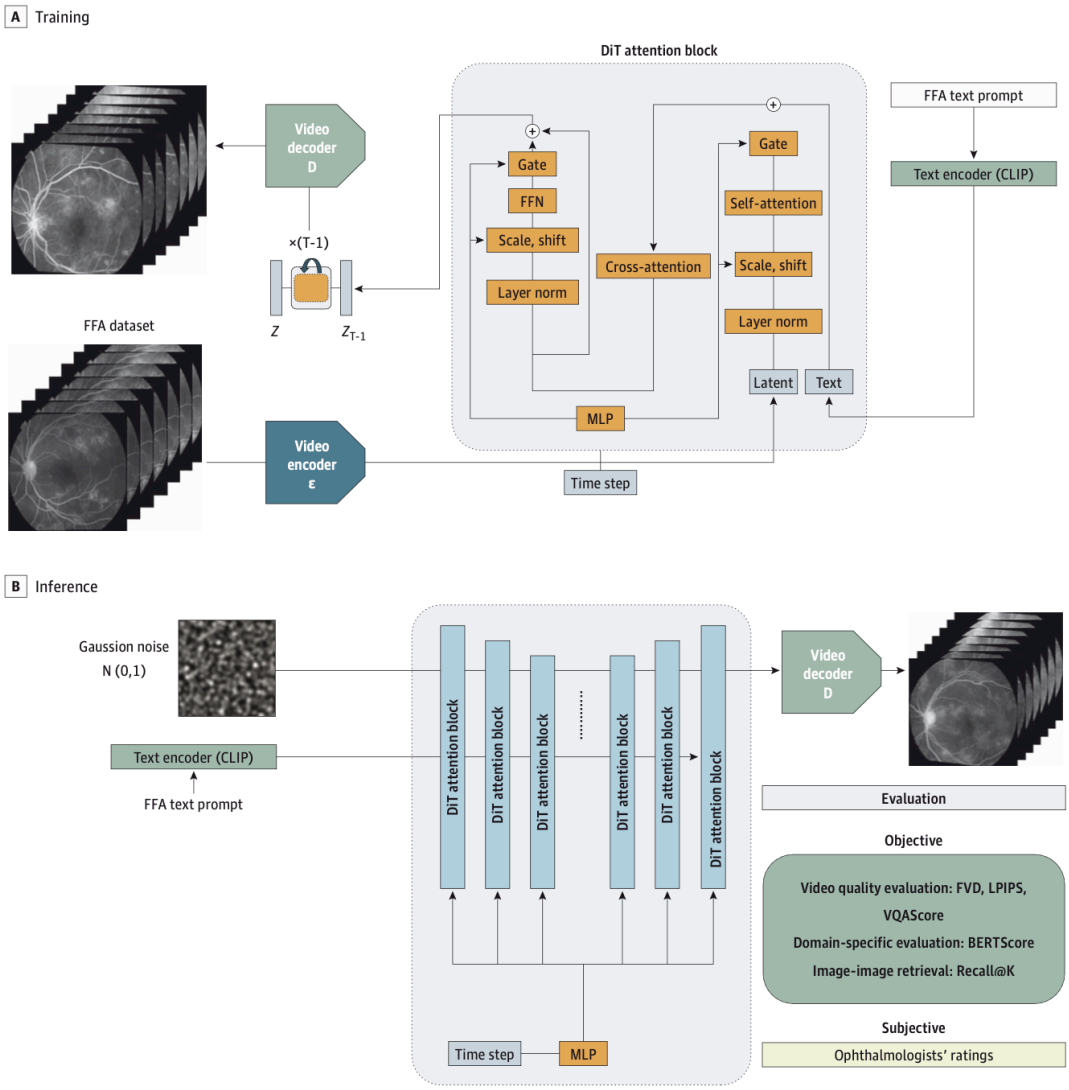


Figure 1. Overview of the text-to-video diffusion model. This figure illustrates a text-to-video framework for generating fundus fluorescein angiography (FFA) sequences using a diffusion transformer (DiT). BERTScore indicates bidirectional encoder representations from transformers score; FFN, feed-forward network; FVD, Fréchet video distance; LPIPS, learned perceptual image patch similarity; MLP, multilayer perceptron; VQAScore, visual question answering score. In A (during training), the input FFA video sequences are first encoded into latent representations by the video encoder ε. These latent representations undergo progressive refinement through multiple DiT attention blocks, which integrate self-attention for capturing spatial-temporal dependencies, cross-

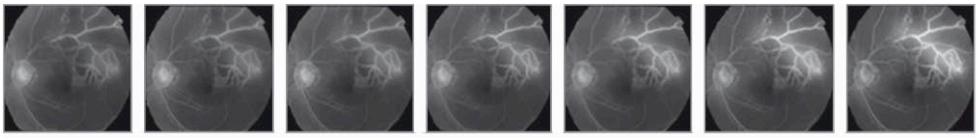
attention for conditioning on textual embeddings, and diffusion-based denoising guided by time-step embeddings. A pretrained text encoder (contrastive language–image pretraining [CLIP]) translates medical text prompts into semantic embeddings, providing textual conditioning that guides the denoising process. The refined latent representations are subsequently decoded by the video decoder D to reconstruct realistic FFA video sequences aligned with the input textual descriptions, potentially enabling clinically meaningful ophthalmic imaging. In B (during inference), the model synthesizes FFA sequences from Gaussian noise conditioned solely on clinical text prompts. The framework is evaluated through objective metrics and subjective evaluations by ophthalmologists to assess visual quality and clinical relevance. The details of the FFA text prompt are as follows: Severe nonproliferative diabetic retinopathy at right eye FFA: (1) The arterial filling and venous return time were basically normal. (2) Multiple microaneurysmal hyperfluorescence can be seen in all quadrants, accompanied by mild leakage. No perfusion defects were formed around the periphery, and obvious vascular leakage was seen in the nonperfusion zone. Diffuse capillary leakage in the posterior pole, obvious leakage around the macula, and no obvious macular edema. (3) No obvious abnormal fluorescence was found in the optic disc.

A Branch retinal vein occlusion

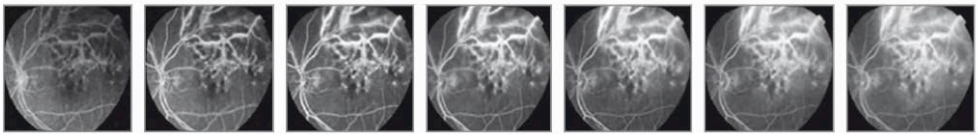
Text prompt

Ischemic branch retinal vein occlusion, nonperfusion area at left eye FFA:
1. Retinal artery diameter is narrowed, with the formation of temporal branch arteriovenous pressure imprint, slow flow of temporal branch venous reflux, and visible venous anastomosis on the temporal side of the macula; small blood vessels in the area governed by the temporal branch vein are dilated and leaky, with scattered microaneurysms, large areas of nonperfusion forming, affecting the temporal upper quadrant of the macula.
2. No obvious abnormalities were observed in the inferior and nasal midperiphery.
3. No obvious abnormalities were observed in the fluorescence of the optic disc.

Real FFA



Generated FFA

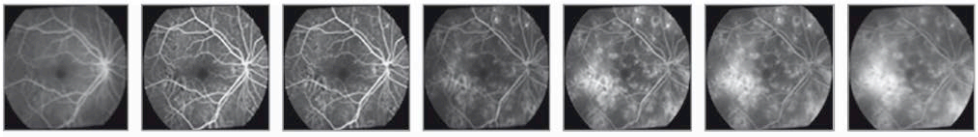


B Severe nonproliferative diabetic retinopathy

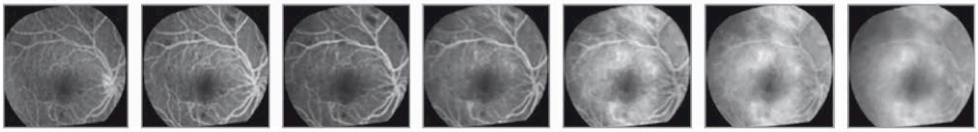
Text prompt

Severe nonproliferative diabetic retinopathy, arteriosclerosis, microaneurysm at right eye FFA:
1. The arterial filling and venous return time were basically normal.
2. Multiple microaneurysmal hyperfluorescence can be seen in all quadrants, accompanied by mild leakage; no perfusion defects were formed around the periphery; and obvious vascular leakage was seen in the nonperfusion zone. Diffuse capillary leakage in the posterior pole, obvious leakage below the temporal macula, and no obvious macular edema.
3. No obvious abnormal fluorescence was found in the optic disc.

Real FFA



Generated FFA

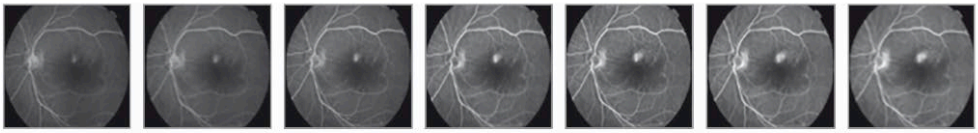


C Age-related macular degeneration

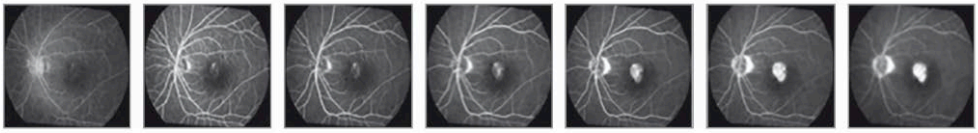
Text prompt

Depigmentation, pigmentary change, hypopigmentation at left eye FFA:
1. Focal choroidal atrophy and pigment epithelial damage with focal leakage of fluorescent dye.
2. No obvious abnormal fluorescence was observed in the midperipheral and peripheral retina.
3. No obvious abnormal fluorescence was observed in the optic disc.

Real FFA



Generated FFA

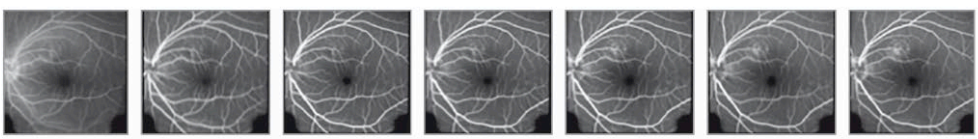


D Central serous chorioretinopathy

Text prompt

Central serous chorioretinopathy, pigmentary change at left eye FFA:
Small early hypofluorescent dots are seen at 1.5 disc diameter temporal to the optic disc, gradually fluorescing with no leakage, and with fluorescein visibility around the lesion. Several small pinpoint areas of hyperfluorescence are seen in the temporal macula. Optic disc appears normal.

Real FFA



Generated FFA

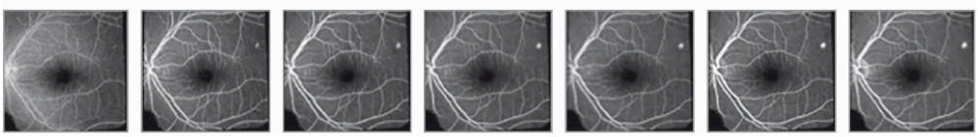


Figure 2. Representative Examples of Real vs Artificial Intelligence–Generated Fundus Fluorescein Angiography (FFA) Videos. BRVO = branch retinal vein occlusion, NPDR = non-proliferative diabetic retinopathy, AMD = age-related macular degeneration, CSC = central serous chorioretinopathy.

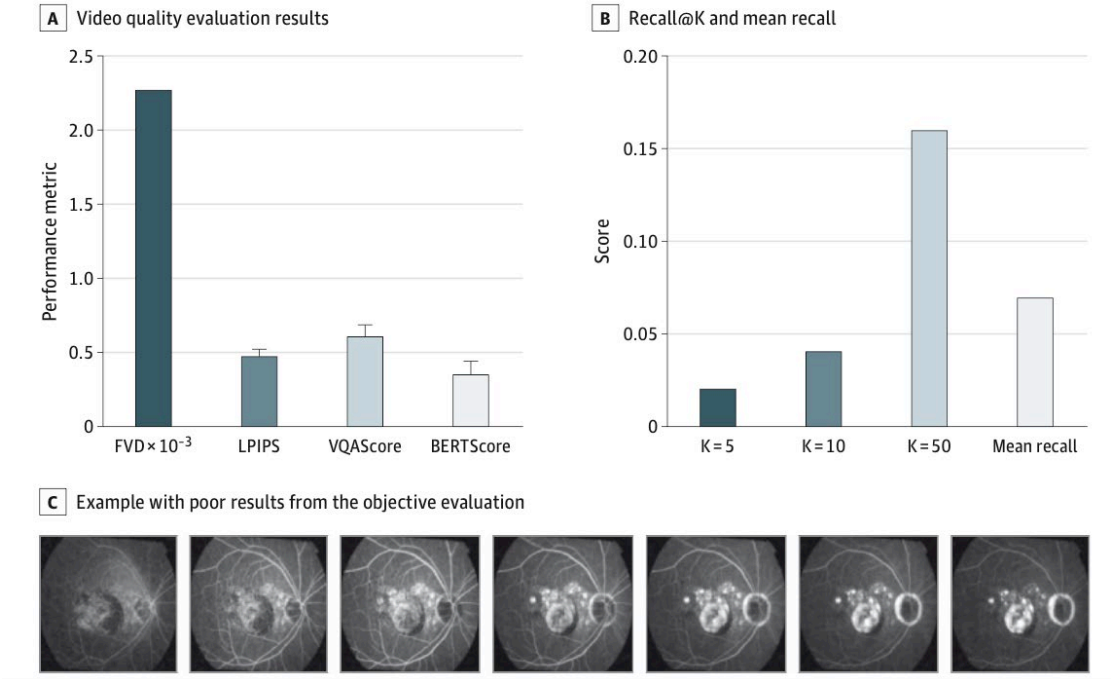


Figure 3. Results From the Objective Evaluations of the Fundus Fluorescein Angiography (FFA) Video Quality. BERTScore indicates bidirectional encoder representations from transformers score; FVD, Fréchet video distance; LPIPS, learned perceptual image patch similarity; VQAScore, visual question answering score. The details of the FFA text prompt are as follows: Depigmentation, pigmentary change, atrophy at right eye FFA: (1) Macular limited retinal pigment epithelial damage, fluorescein leakage and tissue hyperfluorescence staining. Macular limited retinal choroidal atrophy and pigmented spot deposition. Focal loss of retinal pigment epithelium pigmentation and pigmented deposition in the posterior pole. (2) Multifocal retinal pigment

epithelial depigmentation and pigmented deposition in the mid-peripheral and peripheral regions. (3) No obvious abnormal fluorescence was observed in the optic disc. In B, the lower the Recall@K value, the smaller the overlapping part between the artificial intelligence–generated FFA videos and the real clinical videos, demonstrating the better performance in preserving patient privacy. In C, the video quality rankings were poor (scores fell within the lowest 20%) across all parameters: the scores were 0.53 for the LPIPS; 0.52 for the VQAScore; and 0.27 for the BERTScore.

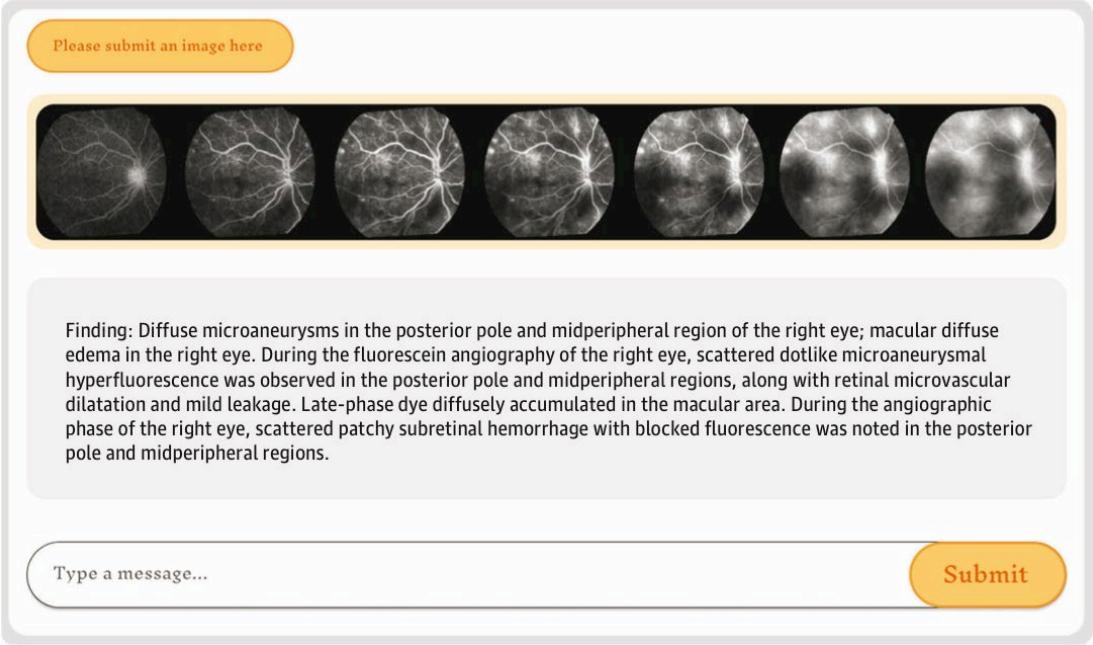
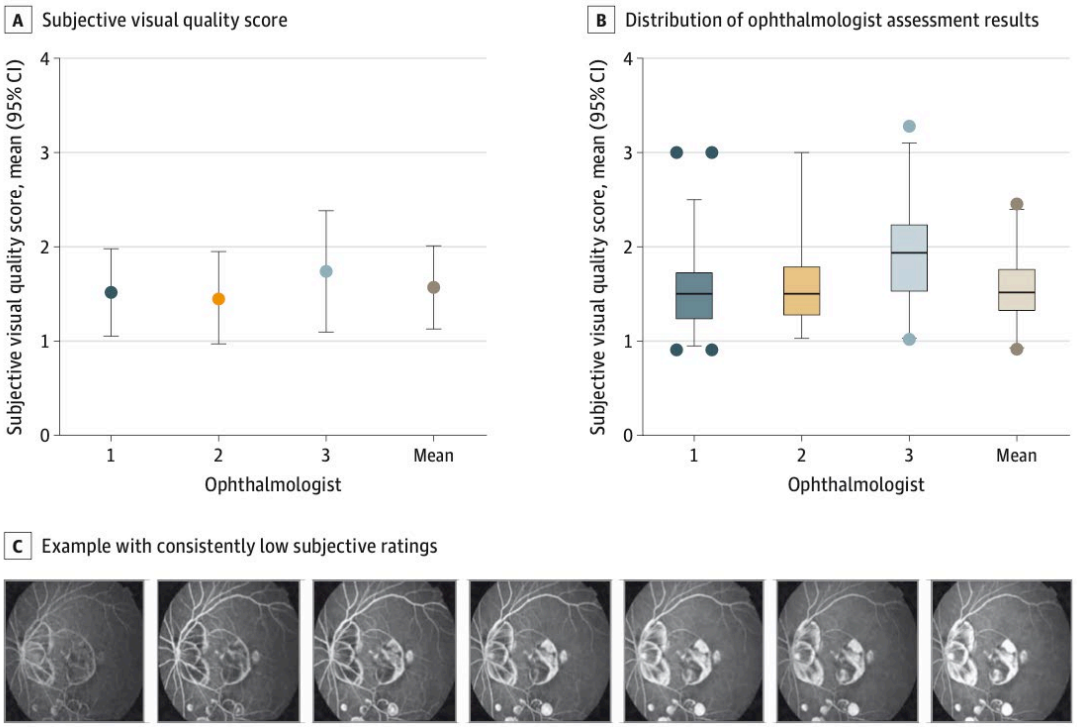


Figure 4. An example of domain-specific evaluation using a Fundus Fluorescein Angiography Generative Pretrained Transformer (FFA-GPT). This figure illustrates an example of the domain-specific evaluation process. An FFA-GPT was used to implement a domain-specific evaluation for determining whether the generated FFA videos accurately depict retinal disease characteristics described in the given prompts



572

573 **Figure 5. Results From the Ophthalmologists' Assessments.** Three ophthalmologists (X.W.,
574 X.Y., and Y.F.) with 3, 5, and 10 years of experience, respectively (representing trainee,
575 attending, and expert levels), reviewed 50 randomly selected FFA videos generated from the
576 test set, comparing them against the corresponding real clinical videos. In A, the 3
577 ophthalmologists assessed video quality on a scale from 1 (excellent) to 5 (very poor). The
578 mean score was 1.52 (SD, 0.46) for the first ophthalmologist; 1.46 (SD, 0.49) for the second
579 ophthalmologist; and 1.74 (SD, 0.65) for the third ophthalmologist. The mean score was 1.57
580 (SD, 0.44) for all 3 ophthalmologists. In B, the first quartile was the 25th percentile of data; the
581 second quartile, the 50th percentile; and the third quartile, the 75th percentile. The whiskers
582 indicate the 2.5th and 97.5th percentiles, representing the range encompassing 95% of the data,
583 and the outliers (shown as filled circles) denote the values outside this range. In C, the artificial
584 intelligence–generated FFA video received consistently low subjective ratings due to its failure
585 to accurately generate the lesions described in the prompt. The details of the FFA text prompt

are as follows: Depigmentation, pigmentary change, atrophy at right eye FFA: (1) Macular limited retinal pigment epithelial damage, fluorescein leakage, and tissue hyperfluorescence staining. Macular limited retinal choroidal atrophy and pigmented spot deposition. Focal loss of retinal pigment epithelium pigmentation and pigmented deposition in the posterior pole. (2) Multifocal retinal pigment epithelial depigmentation and pigmented deposition in the midperipheral and peripheral regions. (3) No obvious abnormal fluorescence was observed in the optic disc.

Supplementary Documents

eFigure 1: Representative FFA videos of other retinal vascular diseases generated by FFA Sora.

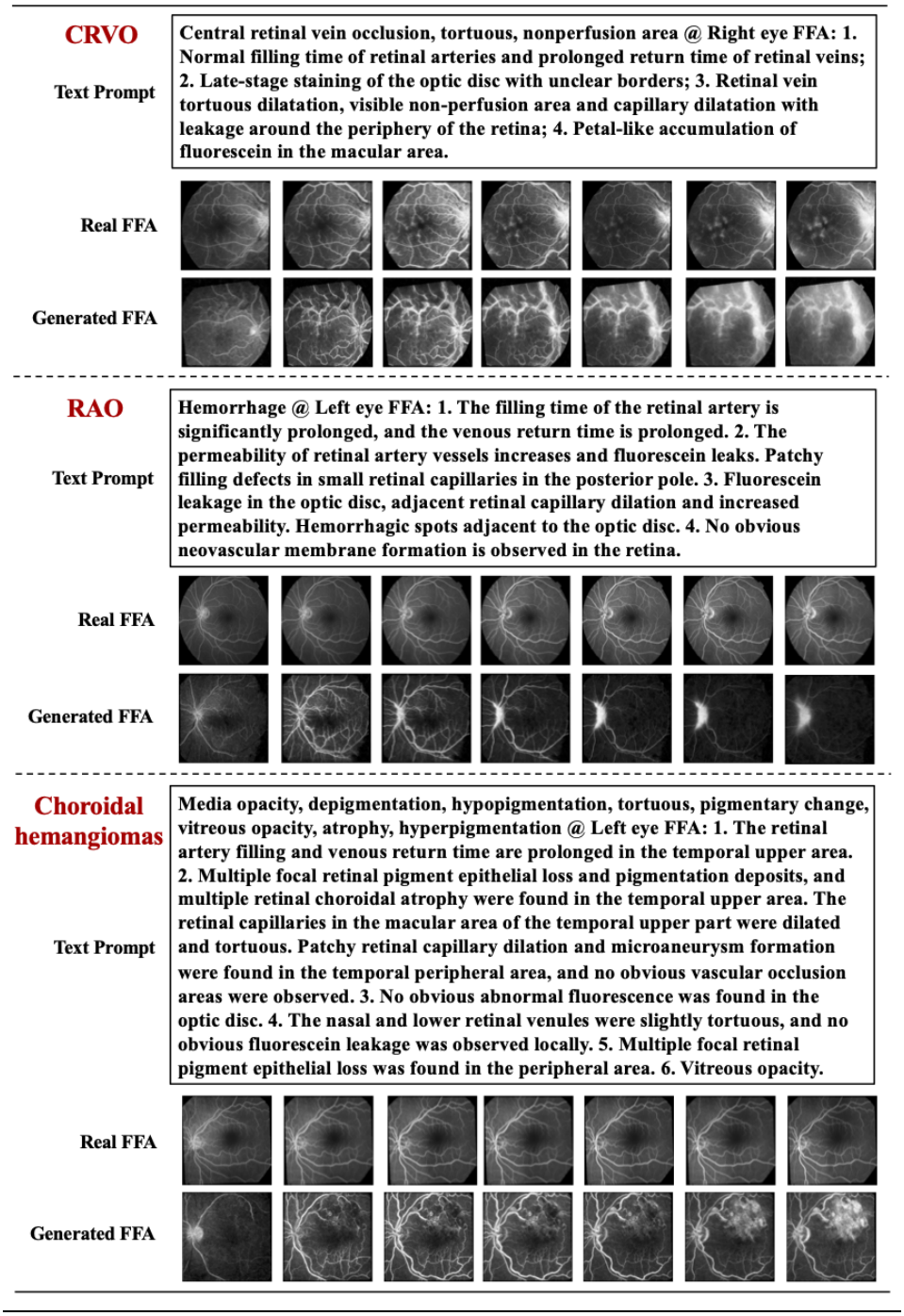
eFigure 2: Representative FFA videos of other RPE-choroidal diseases generated by FFA Sora.

eFigure 3: Examples of **subjective evaluations** across levels (from 1 to 5).

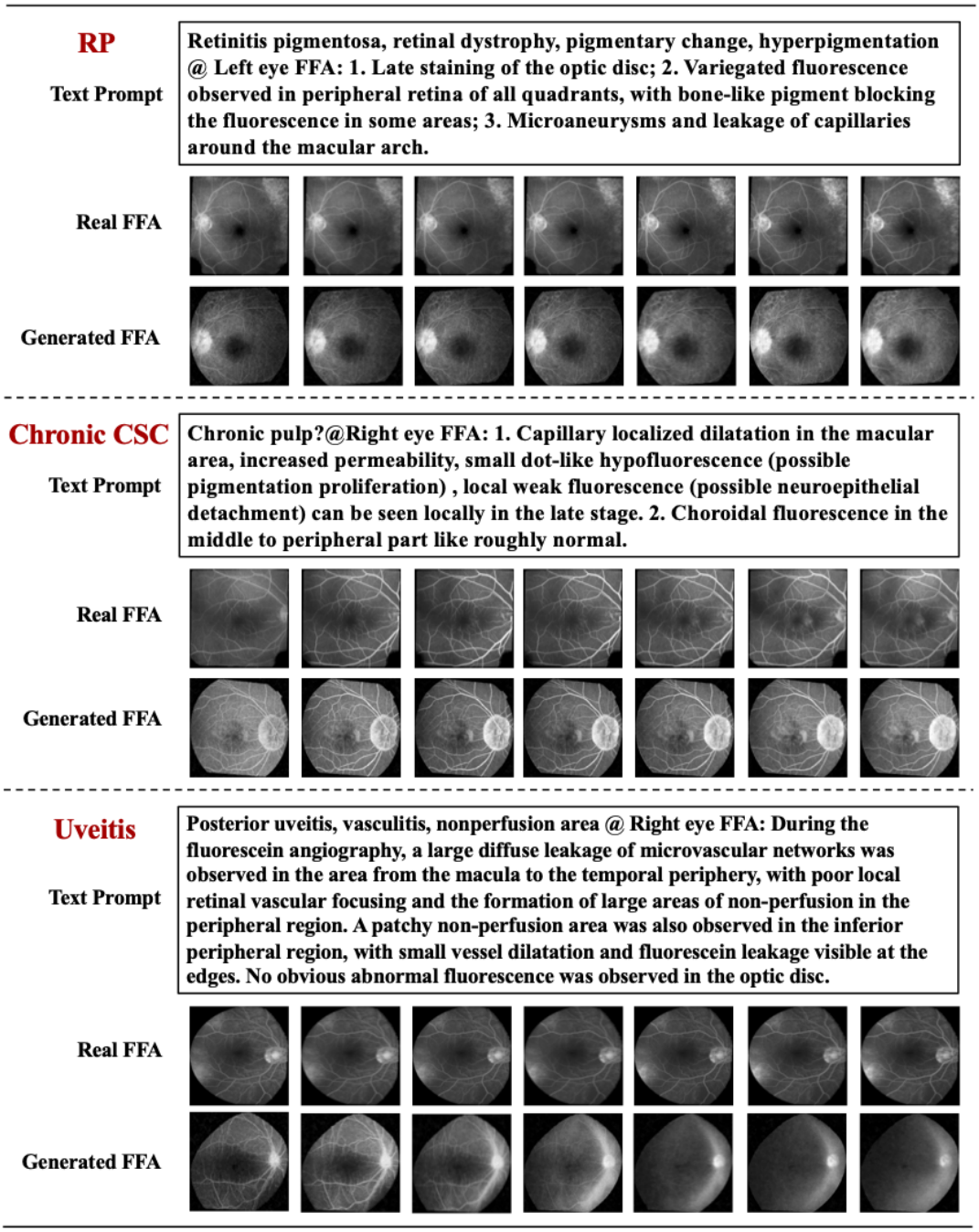
eFigure 4: Examples of inconsistencies between objective and subjective evaluations

eTable 1: Demographic characteristics of the dataset.

eTable 2: The main eye conditions extracted from the fundus fluorescein angiography reports (total N = 3625).

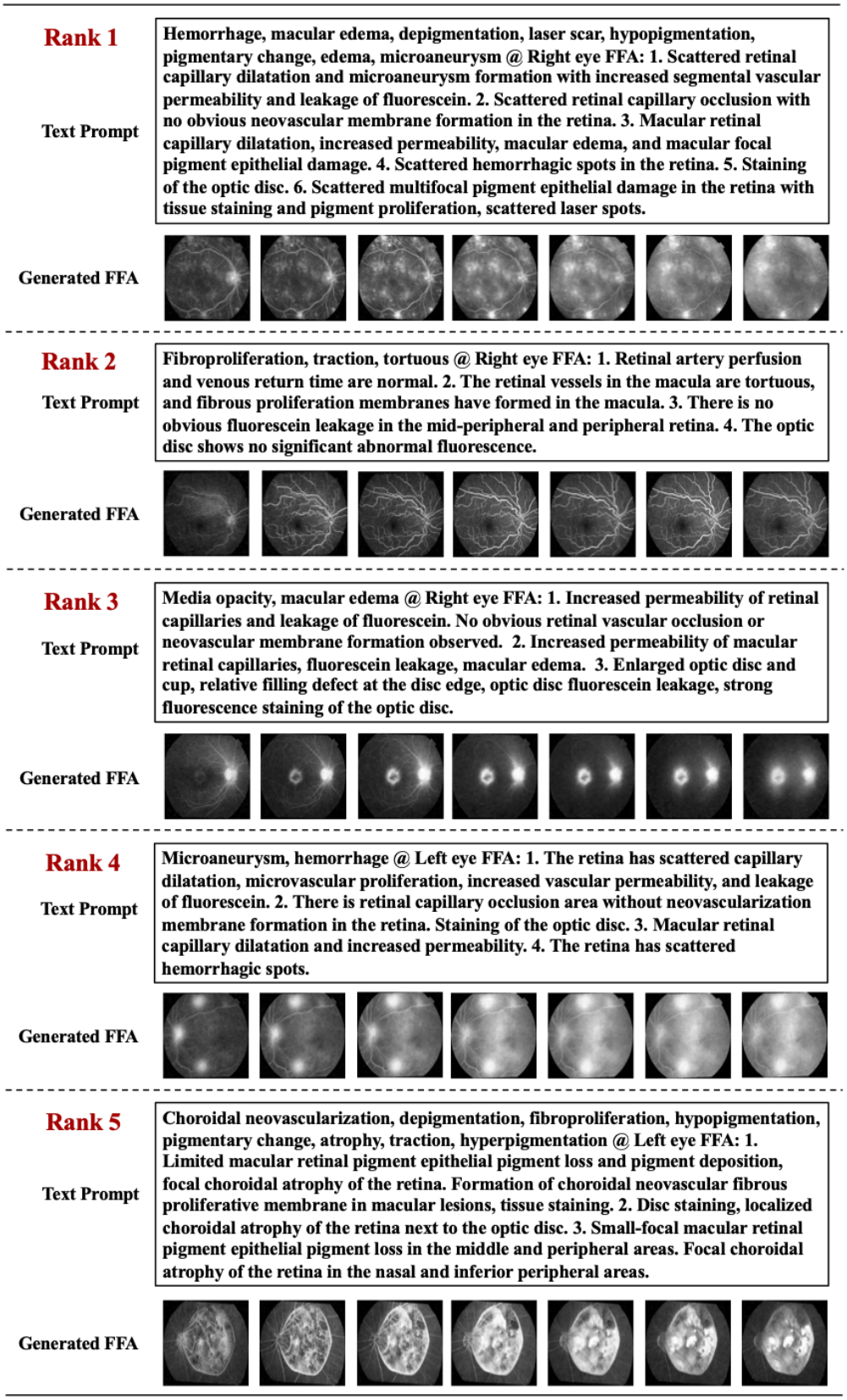


eFigure 1: Representative FFA videos of other retinal vascular diseases generated by FFA Sora. CRVO = central retinal vein occlusion, RAO = retinal artery occlusion.



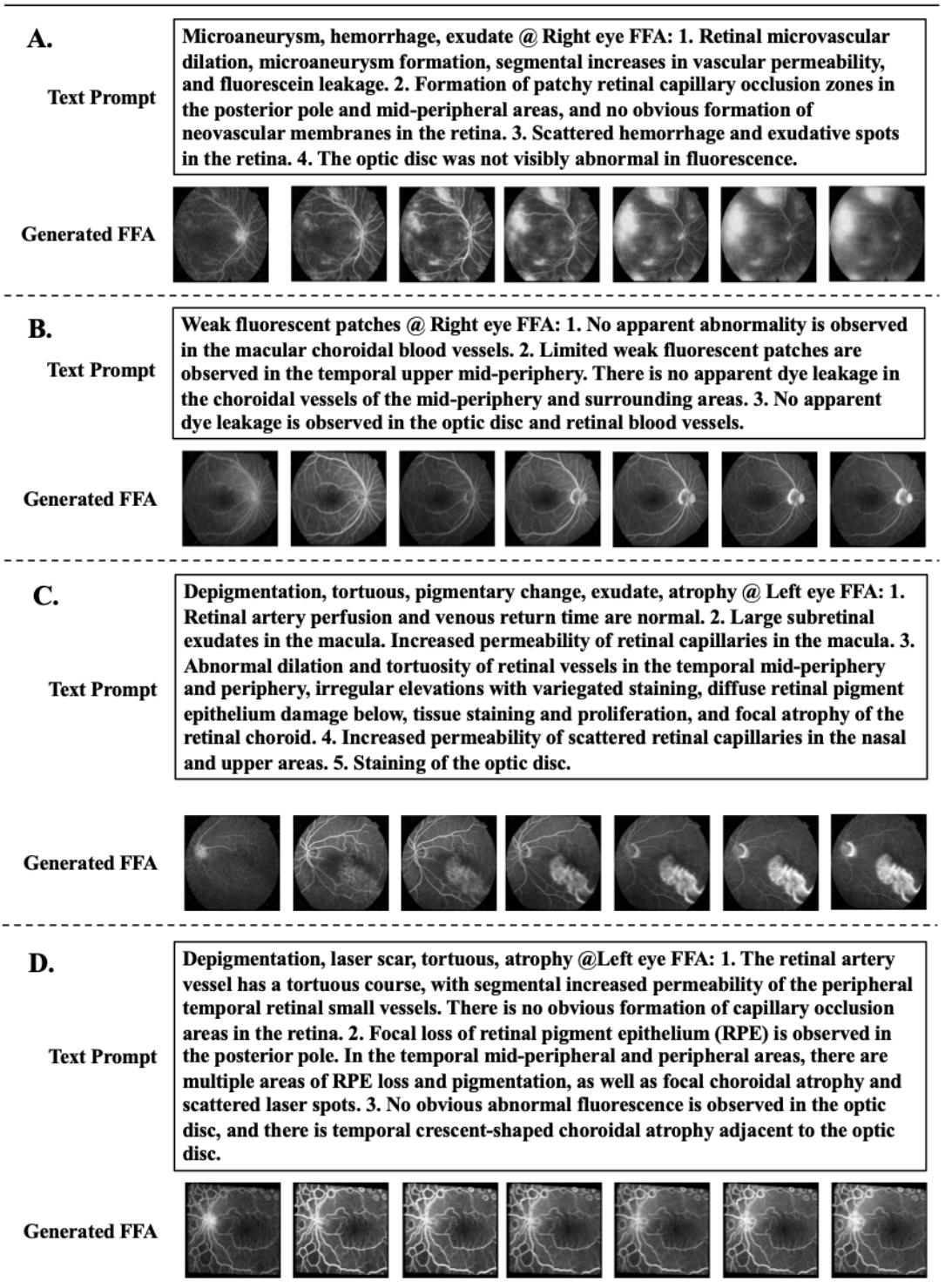
eFigure 2: Representative FFA videos of other RPE-choroidal diseases generated by FFA Sora.

RP = retinal pigmentosa, CSC = central serous chorioretinopathy.



eFigure 2: Examples of subjective evaluations across levels (1 to 5). 1 = The retinal structure and lesion characteristics of the generated videos exactly match the text prompts; 2 = The retinal structure of the generated videos corresponds to the text prompts, and the lesion characteristics

are basically consistent with the text prompts. In this example, fibroproliferation, a key feature in the text description, is missing from the generated images; 3 = The retinal structure of the generated videos corresponds to the text prompt, and the lesion characteristics are slightly consistent with the text prompts. In this example, the fluorescence leakage is overly pronounced, and the optic disc structure appears indistinct, reducing the clarity of anatomical details; 4 = The retinal structure of the generated videos corresponds to the text prompt, but the lesion features cannot be generated. In this example, only microaneurysms are generated, and the fluorescence signal is excessively strong, failing to reflect the hemorrhagic changes described in the prompt; 5 = Unable to generate all text-oriented features.



eFigure 3: Examples of inconsistencies between objective and subjective evaluations. (A-B) Examples where objective evaluation failed but subjective evaluation results were good. (C-D) Examples where objective evaluation results were good but subjective evaluation failed.

eTable 1: Demographic characteristics of the dataset.

	Total	Train	Validation	Test	P value
No.	3625	2851	387	387	
Age, median (lQR)	50 (36, 61)	50 (36, 61)	48 (34, 62)	50 (37, 62)	0.544
Sex, n (%)					0.685
Female	1636 (45.1)	1309 (45.9)	171 (44.2)	183 (47.3)	
Male	1931 (53.3)	1542 (54.1)	216 (55.8)	204 (52.7)	

eTable 2: The main eye conditions extracted from the fundus fluorescein angiography reports (total N = 3625).

Conditions	*N (%)*
Retinitis	2757(30.3%)
Leakage	2010(22.1%)
Neovascularization	720(7.9%)
Atrophy	673(7.4%)
Microaneurysms	670(7.4%)
Edema	654(7.2%)
Hemorrhage	571(6.3%)
Retinopathy	511(5.6%)
Exudates	166(1.8%)
Uveitis	103(1.1%)
Drusen	86(0.9%)
Chorioretinitis	73(0.8%)
Ischemia	38(0.4%)
Scleritis	24(0.3%)
Vein Sheathing	20(0.2%)
Fibrosis	7(0.1%)
Foveal Cyst	3(0.0%)