This is the Pre-Published Version.

This article has been accepted for publication in British journal of ophthalmology, 2024 following peer review, and the Version of Record can be accessed online at https://doi.org/10.1136/bjo-2023-324446.

ICGA-GPT: Report Generation and Question Answering for Indocyanine Green Angiography Images

Xiaolan Chen,[1#] Weiyi Zhang,[1#] Ziwei Zhao,[1] Pusheng Xu,[4] Yingfeng Zheng,[4] Danli Shi,[1,2*] Mingguang He[1,2,3]

[1] School of Optometry, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China.

[2] Research Centre for SHARP Vision (RCSV), The Hong Kong Polytechnic University, Kowloon, Hong Kong, China.

[3] Centre for Eye and Vision Research (CEVR), 17W Hong Kong Science Park, Hong Kong, China.

[4] State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangdong Provincial Clinical Research Center for Ocular Diseases, Guangzhou, Guangdong, China.

[#] Contributed equally

Corresponding Author:

Dr Danli Shi, MD, PhD

The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China

Email: danli.shi@polyu.edu.hk

WHAT IS ALREADY KNOWN ON THIS TOPIC

The interpretation of ophthalmic images requires a significant investment of time and labour. Artificial intelligence technology has achieved automated report generation based on fundus photographs and fluorescein angiography images but not on indocyanine green angiography (ICGA) images.

WHAT THIS STUDY ADDS

In this proof-of-concept study, we successfully developed and validated an innovative ICGA-GPT model related to ICGA images. This model combines a multimodality transformer architecture and a large language model, resulting in satisfactory report generation and interactive question-answering performance across both qualitative and quantitative evaluations.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

The introduction of ICGA-GPT as a valuable tool holds promising prospects in the realm of ocular image interpretation. With further optimization and validation in the future, it has the potential to alleviate the day-to-day workload in ophthalmic clinics.

ABSTRACT

Background Indocyanine green angiography (ICGA) is vital for diagnosing chorioretinal diseases, but its interpretation and patient communication require extensive expertise and time-consuming efforts. We aim to develop a bilingual ICGA report generation and question-answering (QA) system.

Methods Our dataset comprised 213,129 ICGA images from 2,919 participants. The system comprised two stages: image-text alignment for report generation by a multimodal transformer architecture, and large language model (LLM)-based QA with ICGA text reports and human-input questions. Performance was assessed using both qualitative metrics (including Bilingual Evaluation Understudy (BLEU), Consensus-based Image Description Evaluation (CIDEr), Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence (ROUGE-L), Semantic Propositional Image Caption Evaluation (SPICE), accuracy, sensitivity, specificity, precision, and F1 score) and subjective evaluation by three experienced ophthalmologists using 5-point scales (5 refers to high quality).

Results We produced 8,757 ICGA reports covering 39 disease-related conditions after bilingual translation (66.7% English, 33.3% Chinese). The ICGA-GPT model's report generation performance was evaluated with BLEU scores (1-4) of 0.48, 0.44, 0.40, and 0.37, CIDEr of 0.82, ROUGE of 0.41, and SPICE of 0.18. For disease-based metrics, the average specificity, accuracy, precision, sensitivity, and F1 score were 0.98, 0.94, 0.70, 0.68, and 0.64, respectively. Assessing the quality of 50 images (100 reports), three ophthalmologists achieved substantial agreement (kappa=0.723 for completeness, kappa=0.738 for accuracy), yielding scores from 3.20 to 3.55. In an interactive QA scenario involving 100 generated answers, the ophthalmologists provided scores of 4.24, 4.22, and 4.10, displaying good consistency (kappa=0.779).

Conclusion This pioneering study introduces the ICGA-GPT model for report generation and interactive QA for the first time, underscoring the potential of LLMs in assisting with automatic ICGA image interpretation.

Keywords large language model, indocyanine green angiography, medical report generation, medical question answering, generative artificial intelligence

Introduction

Indocyanine green angiography (ICGA) is an advanced imaging technique that utilizes the indocyanine green (ICG) dye to enhance fundus photographic imaging. ICG exhibits a unique capability to effectively traverse through blood, pigments, and exudate, rendering it particularly well-suited for the visualization of choroidal circulation and analysis of diverse chorioretinal disorders, including choroidal neovascularization (CNV), chronic central serous chorioretinopathy and choroidal haemangiomas, facilitating disease diagnosis and management.[1 2]

However, the interpretation of ICGA images requires a significant level of professional expertise and extensive training. Ophthalmologists are not only tasked with accurate image interpretation and reporting based on ICGA images, but they also face the considerable task of explaining these reports to patients. This process involves a significant investment of time and effort.[3 4] Artificial intelligence (AI) offers immense potential in this context, given its ability to process large volumes of data at a scale beyond human capabilities in various medical domains.[5] With advancements in computer vision, nowadays AI can automatically analyze imaging data.[6 7] Previous studies have tackled the automation of ophthalmic disease classification based on ocular images[4 8] and the generation of reports using fundus fluorescein angiography (FFA) images.[9] However, to the best of our knowledge, a free text-based system for generating ICGA reports has yet to be developed.

Furthermore, existing auto-report systems primarily generate text reports that resemble those generated by humans,[9 10] lacking the necessary interactivity for subsequent comprehension. This often results in patients struggling to fully understand their condition after receiving the report, and doctors being unable to allocate excessive time for explanation due to their already demanding schedules. Recent strides in large language models (LLM) offer a new possibility.[11] These models possess powerful natural language processing and generation capabilities, enabling them to generate context-aware responses during interactions with users, potentially better meeting the needs of both patients and doctors.[12] However, it was reported that even the most advanced model, GPT-4V(sion), struggles with tasks related to ophthalmic image interpretation.[13] The potential of extending LLMs to interpret ophthalmic images and address patient’s concerns remains unexplored.

In light of this, we aim to develop a bilingual ICGA-GPT system, providing a scheme that combines the power of LLMs and report generation models for the dual purposes of report generation and interactive question-answering (QA) based on the imaging report.

Methods

The study flow chart is shown in **Figure 1**.

Dataset

The data for this study were retrospectively collected from a tertiary eye hospital in China between November 2016 and December 2019. All patient data were anonymized, adhering to the Declaration of Helsinki's principles. We excluded low-quality ICGA images by extracting vessels,[14] where images with detectable vessel ratios less than 0.01 were excluded. Each case contained an average (SD) of 146.61 (70.01) ICGA images captured using the Heidelberg spectral camera (Heidelberg, Germany) with a resolution of 768 × 768 pixels. We employed a temporal splits strategy to validate our model. It is a temporal or narrow validation that involves developing a model using past data and validating it on future data.[15] Specifically, we divided the dataset into the training set (data prior to June 2019), the validation set (data from June to September 2019), and the testing set (data from September to December 2019, serving as a form of external validation). Anonymous patient IDs were used to ensure there is no patient overlap between different datasets, thus preventing any potential data leakage or result bias arising from cases in previous visits.

Development of ICGA-GPT

We first translated the initial Chinese ICGA report into English to obtain bilingual (Chinese and English) versions of the reports. We employed the Bootstrapping Language-Image Pre-training (BLIP) framework. This framework has been pre-trained on several datasets, including two human-annotated datasets (COCO and Visual Genome), and three web datasets (Conceptual Captions, Conceptual 12M, SBU captions),[16] thus enabling the model to learn visual features and semantic meaning from a diverse and extensive collection of images. It consists of two main components, visual transformer[17] as the image encoder and Bidirectional Encoder Representations from Transformers[18] as the language encoder and decoder. Using the ICGA images and their corresponding bilingual reports, we fine-tuned the pretrained BLIP model. During the training process, we randomly selected 1-9 images from each case as input for the model. The sampling was balanced between the early, mid, and late phases of ICGA. All images were resized to 320x320 pixels. We used the AdamW optimizer during the fine-tuning process. The initial learning rate was set at 0.00002, accompanied by a weight decay of 0.05, and a cosine learning rate schedule. The fine-tuning was conducted with a preset of 50 epochs using two NVIDIA Tesla V100 GPUs, the model with the best Bilingual Evaluation Understudy (BLEU)1 score on the validation set was chosen for testing.

We further introduced the state-of-the-art open-source LLM Llama 2-7b[12] in the second step to explore its capability to read generated reports. Specifically, we selected high-quality English reports covering various conditions and formulated a series of typical questions related to ICGA reports based on our clinical experience per methods similar to those used by Momenaei et al.[19] For each report, we paired it with the prepared questions and inputted them into the Llama 2 model for testing.

Evaluation of ICGA-GPT

Evaluation of the report generation part

Language-based metrics

Our model aims to generate ICGA reports that closely resemble those written by human experts. The quality of generated reports is evaluated using the general quantitative natural language generation (NLG) metrics, including BLEU,[20] Consensus-based Image Description Evaluation (CIDEr),[21] Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence (ROUGE-L),[22] and Semantic Propositional Image Caption Evaluation (SPICE).[23] Considering the number of words in common ophthalmology phrases , we utilize 1-4 grams of BLEU to assess the predicted reports.

Disease-based metrics

To address the shortcomings of the conventional NLG metrics in medical abnormality detection, we further included disease-based evaluation metrics, including accuracy, sensitivity, specificity, precision, and F1 score. Specifically, we employed a keyword-matching approach to extract and standardize the terminology for disease conditions, which formed the foundation for evaluating the generated reports. Subsequently, we identified 14 commonly diagnosed conditions in ophthalmic clinical practices using ICGA. For each of these conditions, we computed individual top-1 classification metrics, as well as the average metrics across all 14 abnormalities.

Human-assessment

A total of 50 images from the test set were randomly sampled, and their corresponding bilingual generated reports (50 in Chinese and 50 in English) were manually assessed by three qualified ophthalmologists (X.C., Z.Z. and P.X.) using a scale with 5-point scores. The evaluation focused on two aspects: (1) completeness: the degree of condition match between ground truth and the generated reports, and (2) correctness: the accuracy of angiographic and condition descriptions. The ground truths were original reports written by professional ICGA specialists, serving as reliable references for evaluation. The scale incorporated five response options, with each option assigned a value ranging from 1 (strongly disagree) to 5 (strongly agree). The details of the scale are presented in online **Supplemental Table S1**. The ophthalmologists first validate the original reports with the ICGA images to ensure their reliability, and then compare them with the generated reports and assign scores to each aspect of each report. The average value of the three evaluators is recorded as the final score for each item. To determine the agreement between the ophthalmologists, we calculated Fleiss' Kappa. The value of Kappa (k) ranges from -1 to 1 and is used as an indicator of the strength of agreement where the Kappa statistics were interpreted as follows: 0.01-0.20 (slight agreement), 0.21-0.40 (fair agreement), 0.41-0.60 (moderate agreement), 0.61-0.80 (substantial agreement), and 0.81-0.99 (almost perfect agreement).[24]

Evaluation of the QA part

Human-assessment

Each set of recorded responses to all the prepared questions was also reviewed by the three ophthalmologists. The ophthalmologists assessed the answers subjectively, considering factors such as inappropriate content, missing content, and extent of possible harm,[25] on a scale of 1 to 5 (5=excellent, 4=good, 3=normal, 2=poor, and 1=very poor), with a score 5 referring to the highest quality. The final score of the answer is obtained by averaging the ratings given by three evaluators. The detailed grading criteria and examples of different quality were demonstrated in online **Supplemental Table S2** and online **Supplemental Figure S1**. The inter-rater agreement was also assessed using the same metric, Fleiss' Kappa, as in the report evaluation.

Results

The final dataset included 213,129 ICGA images alongside 2,919 reports. Among these images, 132,452 (62.1%) were in early-phase (57,668 from 20 to 60 seconds and 74,784 from 60 seconds to 3 minutes), 44,726 (21.0%) were in mid-phase and 35,951 (16.9%) were in late-phase. The median (interquartile range) age of the participants was 54 (40, 65) years, and 1662 (56.9%) were male. Detailed characteristics of the dataset are presented in **Table 1**.

Table 1 Indocyanine green angiography (ICGA) dataset characteristics**.**

	Total	**Train**	**Validation**	**Test**	**P value**
Population					
No.	2919	1612	580	727	
Age, median (IQR)	54 (40, 65)	54 (42, 65)	53 (37, 65)	54 (38, 66)	0.134
Sex, n (%)					0.025
Female	1257 (43.1)	662 (41.1)	253 (43.6)	342 (47.0)	
Male	1662 (56.9)	950 (58.9)	327 (56.4)	385 (53.0)	
Year, n (%)					< 0.001
2016	149 (5.1)	149 (9.2)	0 (0)	0 (0)	
2017	883 (30.3)	883 (54.8)	0 (0)	0 (0)	
2019	1887 (64.6)	580 (36.0)	580 (100)	727 (100)	
ICGA images					
No.	213129	114880	43463	54786	
Phase*, n (%)					< 0.001
Early	57668 (27.1)	31382 (27.2)	11654 (26.7)	14632 (26.6)	
Early1	74784 (35.1)	39724 (34.4)	15107 (34.6)	19953 (36.3)	
Mid	44726 (21.0)	24379 (21.1)	9132 (20.9)	11215 (20.4)	
Late	35951 (16.9)	19395 (16.8)	7570 (17.3)	8986 (16.4)	

*Early: 20s-60s; Early1: 60s-3min; Mid: 3min-5min; Late: > 5min.

IQR, Interquartile Range.

We created a dictionary to extract diagnoses and findings with keyword matching from the original Chinese reports. After data extraction, most of the participants were diagnosed with multiple chorioretinal conditions, including hemorrhage (8.7%), drusen (7.6%), pigmentary change (6.5%), CNV (5.7%), and retinal dystrophy (5.0%). There was a total of 39 conditions, and their distribution is illustrated in **Figure 2A**. We generated a total of 8,757 reports, with 5,838 (66.7%) in English and 2,919 (33.3%) in Chinese. The words in reports primarily encompass positional location and specific angiography-related descriptions. The word cloud of the top-appearing words is provided in **Figure 2B**.

Quantitative model performance

The model demonstrated satisfactory performance in generating ICGA reports, as shown in **Table 2**. In terms of language-based evaluation, the model achieved the following scores: BLEU1=0.48, BLEU2=0.44, BLEU3=0.40, BLEU4=0.37, CIDEr=0.82, ROUGE=0.41, and SPICE=0.18. Regarding disease-based metrics, the overall performance for the 14 common diseases was acceptable. The overall specificity, accuracy, precision, sensitivity, and F1 scores were 0.98, 0.94, 0.70, 0.68, and 0.64, respectively. The metrics for each condition are presented in **Table 2B**. In the context of imbalanced distribution, the F1 score is an unbiased metric and is of significant value, with the highest score observed for wet AMD at 0.80.

Additionally, we investigated the relative importance of different numbers of input images to the ICGA-GPT by inputting a range of one to twelve images. We observed that the performance of the model improved dramatically when more than one image was provided and plateaued when there were at least four images (online **Supplemental Figure S2**, pink line) for the BLEU-1 metric. This trend was preserved for CIDEr, ROUGE, and SPICE metrics (online **Supplemental Figure S2**, blue, green, and orange lines).

Table 2 Quantitative model performance in the test set: (A)language-based metrics, (B) disease-based metrics

A.

BLEU_1	BLEU_2	BLEU_3	BLEU_4	CIDEr	ROUGE	SPICE
0.48	0.44	0.40	0.37	0.82	0.41	0.18

B.

Condition	Specificity	Accuracy	Precision	Sensitivity	F1 score
Polypoidal choroidal vasculopathy	0.97	0.95	0.68	0.72	0.70
Choroidal neovascularization	0.95	0.89	0.63	0.75	0.68
Wet age-related macular degeneration	1.00	1.00	1.00	0.67	0.80
Age-related macular degeneration	0.97	0.95	0.72	0.73	0.72
Myopia	1.00	0.99	0.67	0.82	0.74
Choroidal hemangioma	1.00	1.00	0.67	0.67	0.67
Vogt-Koyanagi-Harada disease	1.00	1.00	0.57	1.00	0.73
Multiple evanescent white dot syndrome	1.00	1.00	1.00	0.50	0.67
Aneurysm	1.00	0.99	0.60	1.00	0.75
Choroidal mass	0.99	0.99	1.00	0.44	0.62
Chronic central serous chorioretinopathy	1.00	1.00	1.00	0.33	0.50
Lacquer crack	0.98	0.96	0.55	0.55	0.55
Drusen	0.90	0.80	0.41	0.53	0.46
Pathologic myopia	1.00	0.98	0.25	0.80	0.38
Average	**0.98**	**0.94**	**0.70**	**0.68**	**0.64**

BLEU, Bilingual Evaluation Understudy; CIDEr, Consensus-based Image Description Evaluation; ROUGE, Recall-Oriented Understudy for Gisting Evaluation; SPICE, Semantic Propositional Image Caption Evaluation.

219

Qualitative model performance

Qualitative model performance results are provided in **Table 3**. Example-generated reports and answers are shown in online **Supplemental Figure S3**.

Table 3 Qualitative model performance in the test set: (A) report generation, (B) medical question answering

	Rater 1 Mean (SD)	**Rater 2 Mean (SD)**	**Rater 3 Mean (SD)**	**Kappa**
A.				
Completeness	3.55 (1.34)	3.41 (1.38)	3.55 (1.34)	0.723
Accuracy	3.34 (1.36)	3.20 (1.38)	3.35 (1.40)	0.738
B.				
Score	4.24 (1.13)	4.22 (1.18)	4.10 (1.19)	0.779

SD, Standard Deviation.

Quality of generated reports

The mean (standard deviation [SD]) of the scores was 3.55(1.34) for completeness, 3.34(1.36) for accuracy, assessed by the first grader, 3.41(1.38), 3.20(1.38) by the second grader, and 3.55(1.34), 3.35(1.40) by the third grader. Fleiss' Kappa values indicate a good agreement between the three graders for assessing report quality, with a kappa value of 0.723 for completeness and 0.738 for accuracy, respectively.

Further analysis of the high-quality reports (scoring≥4 points) revealed that these reports primarily pertain to conditions such as CNV, drusen, PCV, and age-related macular degeneration, as well as reports with negative findings. However, a small percentage of reports were of poor quality (12.7% scored 1 for completeness and 13% scored 1 for accuracy), predominantly comprising rare conditions such as retinal folds and Behcet's disease, as well as conditions like media opacity that can result in blurry images.

Quality of question-answering

The question lists we created contain a total of 20 items, covering a variety of topics, including report summaries, disease definitions, etiology, the visual impact of the condition, preventative measures, further examinations, treatment options, prognosis, related complications, and ICGA testing information including timing, specific phases, and post-test instructions (See online **Supplemental Table S3**). We each selected a representative high-quality English report of CNV, drusen, PCV, age-related Macular degeneration, and lacquer crack, respectively, and input a series of prepared questions. The results showed that the mean (SD) of the answer scores was 4.24(1.13) by the first grader, 4.22(1.18) by the second grader, and 4.10(1.19) by the third grader. The kappa value for answer scores was 0.779, indicating a substantial level of agreement among the three graders. Among all the answers related to different diseases, high-quality (⩾4 points) answers accounted for 70% in AMD, 60% in CNV, 80% in drusen, 90% in lacquer crack, and 75% in PCV.

Discussion

In this study, we employed a multi-modal transformer to bridge the gap between ICGA images and reports, and then utilized an LLM to achieve interactive question-answering. The system exhibited consistent and satisfactory performance, as evaluated through both quantitative and qualitative metrics. To the best of our knowledge, this represents the first attempt to showcase the potential of LLMs in assisting with report generation and subsequent QA to enhance the interpretation of ICGA images.

Traditional image captioning models typically use Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) architecture to translate images into text.[26] However, these methods have limitations on the length of generated text and adaptability to multi-modal data. In contrast, Transformers have an advantage in these issues due to their cross-attention architectures.[10 27] In the realm of ophthalmic imaging, transformer-based models have been used in image-text conversion tasks. Wu et al.[28] developed a hybrid CNN-Transformer system for retinal images. Li et al.[9] developed a cross-modal clinical graph Transformer for FFA report generation. A study also used a contrastive pre-training method to enhance medical report generation systems.[29] In our study, we adopted the BLIP,[16] a pure Transformer architecture pre-trained on a large image-text dataset. This demonstrates the effectiveness of adapting pre-trained models to the specific domain of ICGA images. In terms of AI in ICGA, a previous study demonstrated promising classification performance for

PCV,[4] but it was solely focused on the classification task of a specific disease (Specificity=0.80, Accuracy=0.83, Precision=0.81, and Sensitivity=0.87). By finetuning BLIP using a bilingual ICGA dataset, we not only achieved competitive classification performance for PCV (Specificity=0.97, Accuracy=0.95, Precision=0.68, and Sensitivity=0.72) but also enabled discrimination and generation of free-text ICGA-related reports for various chorioretinal diseases. The model's performance in these cases underscores its potential value in clinical settings.

LLMs are powerful forms of AI with world knowledge and logical reasoning abilities.[11] Fine-tuning is a common strategy for using LLMs to solve medical issues but comes with certain limitations, including potential performance reduction in other un-finetuned domains,[30] the need for significant computational resources,[31] and the lack of high-quality medical dialogue data.[25] To fully leverage the in-context learning and reasoning capabilities of LLMs without fine-tuning, we have integrated prompts to create an ICGA report-specific medical dialogue pipeline. LLM serves as the “brain” of the pipeline, analyzing queries to determine if it can provide answers based on the given information. This pipeline is similar to the research conducted by Wang et al,[32] which presented a method of integrating LLMs into computer-aided diagnosis networks for X-ray images. This approach has several advantages. Firstly, while LLMs can produce hallucinations, it is unpredictable when patients input reports to any LLM and receive answers. Our experiments showed that the input prompts matter, and we restricted the LLM to answer questions based on the imaging report. This solution has been professionally evaluated by ophthalmologists, demonstrating its potential in interpreting highly specialized ophthalmic reports, thereby reducing dependence on other unreliable online resources. Secondly, the pipeline offers a user-friendly and comprehensive solution for patients, promoting their active involvement in the decision-making process and potentially enhancing patient autonomy if integrated into a clinical system in the future. Thirdly, Llama-2 is an open-source model with excellent scalability, providing a foundation for incorporating external knowledge and fine-tuning it with reliable domain-specific databases in the future. It can be continuously improved during application, facilitating the delivery of personalized and effective healthcare services to patients.

The utilization of AI in healthcare systems is crucial and necessary due to its ability to enhance precision and accuracy while reducing the waiting time and workload.[33] Firstly, our ICGA-GPT model can serve as a vital auxiliary tool for ophthalmologists. Generating reports for ICGA images is a time-consuming task and diagnostic errors resulting from difficulty and fatigue may occur, particularly in areas where ophthalmologists are unavailable or lack diagnostic expertise. Previous studies have shown that AI assistance is beneficial, which can reduce diagnostic time[34] and improve radiologists' performance.[35] According to the analysis of both the original and generated report in our study, ICGA-GPT can provide additional information beyond visual observation. This can offer supplementary insights for ophthalmologists and free up their time for more complex cases. In addition, our model can also serve as an educational tool for patients. In overloaded ophthalmic clinics, patients may not have enough time to fully understand their examination reports and medical conditions. In this study, ICGA-GPT has demonstrated its ability to provide patients with basic clinical explanations and recommendations related to etiology, abnormalities, treatment, and follow-up. This suggests its immense potential to reduce medical consultation costs and enhance the viability of online medical services.

There are some limitations in this study. Firstly, we did not have access to a completely external dataset for validation. Instead, we adopted a temporal data-splitting approach. This approach simulates the process of developing a system using retrospective data and subsequently validating that system using data collected from the same sites in the subsequent years. Secondly, our model exhibits subpar performance in generating reports for some relatively rare cases and those with media opacity, and occasionally, the generated answers may be overly vague or inaccurate. Therefore, future work could focus on expanding the dataset to include more challenging-to-distinguish conditions, enhancing the LLM through fine-tuning, or equipping it with ophthalmic domain knowledge. Finally, it should be emphasized that our model is currently limited to ICGA images, and further research should focus on developing multimodal imaging models or combining images with electronic medical record data.

In conclusion, this study highlights the efficacy and potential of combining the report-generation models and LLMs for ICGA report generation and subsequent QA. ICGA-GPT is a promising adjunct to enhance medical image interpretation, benefiting both healthcare professionals and patients.

Author Contributions: DS and XC conceived the study. DS and WZ built the deep learning model. XC, WZ, ZZ and PX performed the data analysis. XC wrote the manuscript. All authors have commented on the manuscript. DS is the guarantor.

Conflict of Interest Disclosures: None reported.

Funding/Support: This study was supported by the Start-up Fund for RAPs under the Strategic Hiring Scheme (P0048623) and the Global STEM Professorship Scheme (P0046113) from HKSAR.

Ethics approval: The study was approved by the Ethics Committee (Approved ID: 2021KYPJ164-3). Individual consent for retrospective analysis was waived.

Data availability statement: We do not have permission to redistribute the dataset.

References

1. Yannuzzi LA. Indocyanine green angiography: a perspective on use in the clinical setting. American journal of ophthalmology 2011;**151**(5):745-51. e1

2. 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;**7**(1):34

3. Sen P, Manayath G, Shroff D, Salloju V, Dhar P. Polypoidal choroidal vasculopathy: an update on diagnosis and treatment. Clinical Ophthalmology 2023:53-70

4. Yang J, Zhang C, Wang E, Chen Y, Yu W. Utility of a public-available artificial intelligence in diagnosis of polypoidal choroidal vasculopathy. Graefe's Archive for Clinical and Experimental Ophthalmology 2020;**258**:17-21

5. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. Nature medicine 2022;**28**(1):31-38

6. GPT-4V(ision) System Card. https://openai.com/research/gpt-4v-system-card

7. Tu T, Azizi S, Driess D, et al. Towards generalist biomedical ai. NEJM AI 2024;**1**(3):AIoa2300138

8. Cui T, Lin D, Yu S, et al. Deep Learning Performance of Ultra-Widefield Fundus Imaging for Screening Retinal Lesions in Rural Locales. JAMA ophthalmology 2023;**141**(11):1045-51

9. Li M, Cai W, Verspoor K, Pan S, Liang X, Chang X. Cross-modal clinical graph transformer for ophthalmic report generation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022:20656-65

10. Chen Z, Shen Y, Song Y, Wan X. Cross-modal memory networks for radiology report generation. arXiv preprint arXiv:2204.13258 2022

11. Zhao WX, Zhou K, Li J, et al. A survey of large language models. arXiv preprint arXiv:2303.18223 2023

12. Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 2023

13. Xu P, Chen X, Zhao Z, et al. Evaluation of a digital ophthalmologist app built by GPT4-V (ision). medRxiv 2023:2023.11. 27.23299056

14. Shi D, Lin Z, Wang W, et al. A deep learning system for fully automated retinal vessel measurement in high throughput image analysis. Frontiers in Cardiovascular Medicine 2022;**9**:823436

15. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Annals of internal medicine 2015;**162**(1):55-63

16. Li J, Li D, Xiong C, Hoi S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. International conference on machine learning 2022:12888-900

17. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words:

Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 2020

18. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 2018
19. Momenaei B, Wakabayashi T, Shahlaee A, et al. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. Ophthalmology Retina 2023;**7**(10):862-68
20. Papineni K, Roukos S, Ward T, Zhu W-J. Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th annual meeting of the Association for Computational Linguistics 2002:311-18
21. Vedantam R, Lawrence Zitnick C, Parikh D. Cider: Consensus-based image description evaluation. Proceedings of the IEEE conference on computer vision and pattern recognition 2015:4566-75
22. Lin C-Y. Rouge: A package for automatic evaluation of summaries. Text summarization branches out 2004:74-81
23. Anderson P, Fernando B, Johnson M, Gould S. SPICE: Semantic Propositional Image Caption Evaluation. 2016
24. Mandrekar JN. Measures of interrater agreement. Journal of Thoracic Oncology 2011;**6**(1):6-7
25. Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023;**620**(7972):172-80
26. Oluwasammi A, Aftab MU, Qin Z, et al. Features to text: a comprehensive survey of deep learning on semantic segmentation and image captioning. Complexity 2021;**2021**:1-19
27. Shamshad F, Khan S, Zamir SW, et al. Transformers in medical imaging: A survey. Medical Image Analysis 2023:102802
28. Wu T-W, Huang J-H, Lin J, Worring M. Expert-defined keywords improve interpretability of retinal image captioning. Proceedings of the IEEE/CVF winter conference on applications of computer vision 2023:1859-68
29. Lin Z, Zhang D, Shi D, et al. Contrastive pre-training and linear interaction attention-based transformer for universal medical reports generation. Journal of Biomedical Informatics 2023;**138**:104281
30. Mosbach M, Pimentel T, Ravfogel S, Klakow D, Elazar Y. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. arXiv preprint arXiv:2305.16938 2023
31. Si C, Gan Z, Yang Z, et al. Prompting gpt-3 to be reliable. arXiv preprint arXiv:2210.09150 2022
32. Wang S, Zhao Z, Ouyang X, Wang Q, Shen D. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. arXiv preprint arXiv:2302.07257 2023
33. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its

applications, advantages, limitations, future prospects, and ethical considerations. Frontiers in artificial intelligence 2023;**6**:1169595

34. Tong W-J, Wu S-H, Cheng M-Q, et al. Integration of artificial intelligence decision aids to reduce workload and enhance efficiency in thyroid nodule management. JAMA Network Open 2023;**6**(5):e2313674-e74

35. Lee JH, Hong H, Nam G, Hwang EJ, Park CM. Effect of human-AI interaction on detection of malignant lung nodules on chest radiographs. Radiology 2023;**307**(5):e222976

Figure Legends：

Figure 1. Flow diagram of this study. ICGA, indocyanine green angiography; LLM=large language model.

Figure 2. (A) The distribution of eye conditions extracted from the indocyanine green angiography reports. (B) The English and Chinese reports’ word clouds.

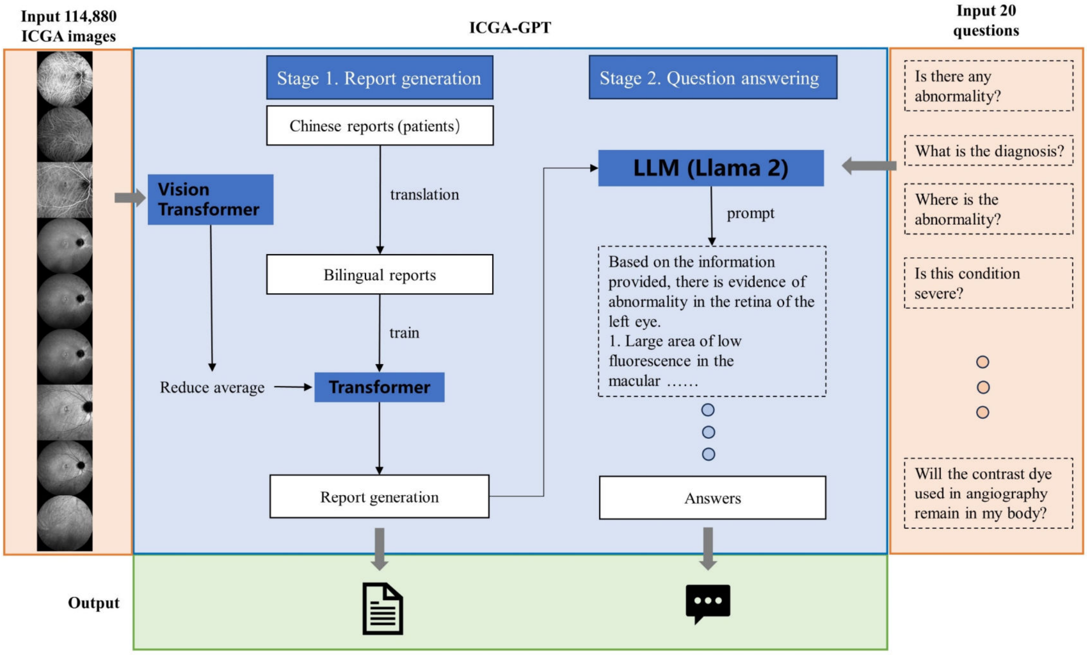

Input 114,880 ICGA images
ICGA-GPT
Input 20 questions
Stage 1. Report generation
Stage 2. Question answering
Chinese reports (patients)
translation
Bilingual reports
train
Vision Transformer
Reduce average
Transformer
Report generation
LLM (Llama 2)
prompt
Based on the information provided, there is evidence of abnormality in the retina of the left eye.
1. Large area of low fluorescence in the macular ……
Answers
Is there any abnormality?
What is the diagnosis?
Where is the abnormality?
Is this condition severe?
Will the contrast dye used in angiography remain in my body?
Output

Figure 1. Flow diagram of this study. ICGA, indocyanine green angiography; LLM=large language model.

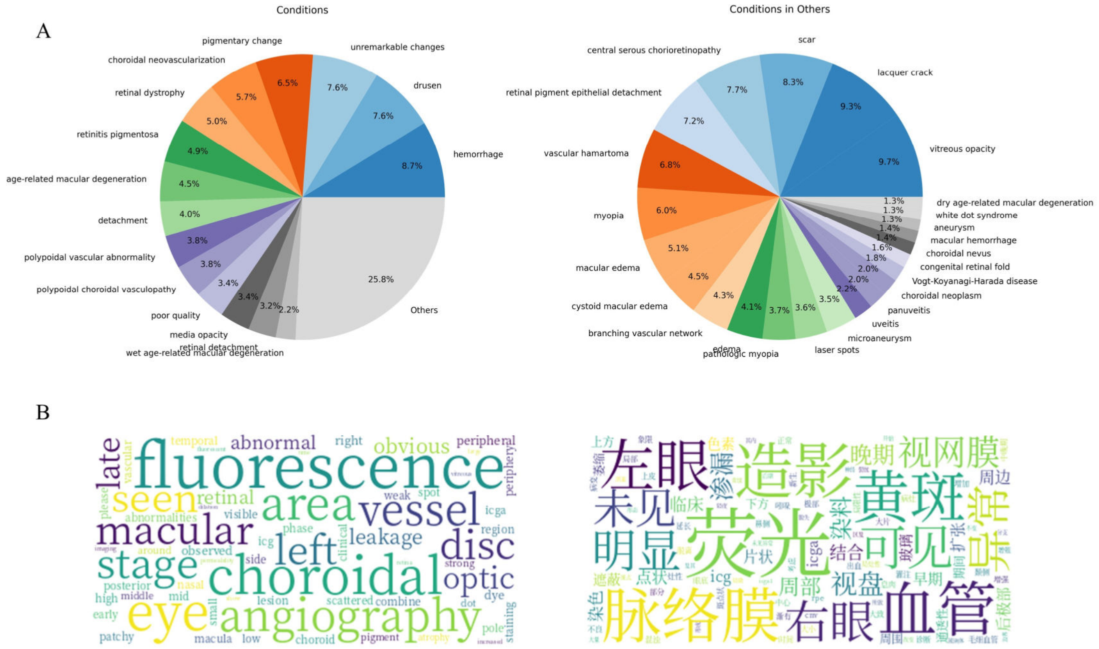


Figure 2. (A) The distribution of eye conditions extracted from the indocyanine green angiography reports. (B) The English and Chinese reports' word clouds.

Supplemental table S1. The scales for qualitative evaluation of report generation.

Report Quality

1. (Completeness) The generated report covers all the lesions in the original images and reports.

Strongly disagree	Disagree	Neither	Agree	Strongly agree
☐1	☐2	☐3	☐4	☐5

2. (Accuracy) The generated report accurately describes the anatomical location and fluorescence characteristics of the lesions.

Strongly disagree	Disagree	Neither	Agree	Strongly agree
☐1	☐2	☐3	☐4	☐5

Supplemental table S2. The scoring criteria for qualitative evaluation of question answering.

Score	**Quality**	**Description**
5	Excellent	◎The answer is essentially correct and can be recommended to patients.
4	Good	◎The answer has a few errors, but it can still be recommended to patients.
3	Normal	◎The answer lacks crucial information and is not suitable for recommending to patients.
2	Poor	◎The answer is incorrect and may bring harm to patients.
1	Very poor	◎The answer is completely wrong or completely unrelated.

Supplemental table S3. The question lists related to ICGA reports for question answering testing.

1. Is there any abnormality?
2. What is the diagnosis?
3. Where is the abnormality?
4. Is this condition severe?
5. Can this condition cause blindness?
6. Will this condition progress?
7. What is the outcome of this condition?
8. What is the prognosis for this condition?
9. What might be the cause of this condition?
10. What are some common complications that can occur?
11. What should I do?
12. What is your suggestion for me?
13. What kind of further testing might be necessary?
14. What might be the treatment options for this condition?
15. How can I prevent this condition?
16. Can you explain how ICGA works and what it can detect?
17. What is the early/mid/late phase of ICGA?
18. How long does an ICGA test usually take to complete?
19. Is there anything I should avoid doing after an ICGA test?
20. Will the contrast dye used in angiography remain in my body?

Supplemental figure S1. Examples of generated answers in different grades. A. high quality. B. medium quality. C. poor quality. Red Text: Marks incorrect information in the model's response.

A

Report **Finding:** 1.There is a large area of choroidal vascular dilation in the macular area, with a large area of low fluorescence in the macular area and increased permeability in the late stage of angiography. 2.There is no obvious abnormalities in the choroidal vessels in the mid-peripheral and peripheral choroidal vessels in the late stage of angiography. 3.No obvious abnormalities were observed in the optic disc and retinal vessels. **Impression:** drusen, detachment, Pigment Epithelial Detachment (PED)	
Score 5	**Question:** What is your suggestion for me?
	Answer: Based on the information you provided, it sounds like you have a condition called choroidal neovascularization (CNV) in the macular area of your retina, which is causing a large area of low fluorescence and increased permeability in the late stage of angiography. This condition is typically associated with age-related macular degeneration (AMD). Here are my suggestions for you: 1. Follow-up monitoring: It is important to have regular follow-up appointments with your retina specialist to monitor the progression of the disease and adjust your treatment plan as needed. 2. Anti-vascular endothelial growth factor (VEGF) injections: These medications can help to reduce the growth of new blood vessels and slow down the progression of the disease. Your retina specialist may recommend these injections as part of your treatment plan. 3. Photodynamic therapy (PDT): This is a treatment that uses a light-sensitive medication and a low-intensity laser to damage and shrink the abnormal blood vessels in the retina. PDT may be recommended as an additional treatment option for you. 4. Lifestyle modifications: Maintaining a healthy diet, exercising regularly, and not smoking can help to slow down the progression of AMD. 5. Vitamin supplements: Certain vitamins, such as vitamin C, E, and beta-carotene, have been shown to slow down the progression of AMD. Your retina specialist may recommend these supplements as part of your treatment plan. It is important to note that the best course of treatment for you will depend on the specifics of your condition and your individual needs. Your retina specialist will be able to provide personalized recommendations based on your specific situation.
Score 4	**Question:** Will this condition progress?
	Answer: Based on the information provided, the condition is likely to progress. The presence of a large area of choroidal vascular dilation in the macular area, low fluorescence in the macular area, and increased permeability in the late stage of angiography are all indicative of choroidal neovascularization (CNV), which is a common complication of age-related macular degeneration (AMD). The fact that there are no obvious abnormalities in the mid-peripheral and peripheral choroidal vessels suggests that the CNV is localized to the macular area. The late stage of angiography and the presence of increased permeability also suggest that the CNV is in an advanced stage. The lack of obvious abnormalities in the optic disc and retinal vessels suggests that the CNV is not causing significant damage to these structures. Given the advanced stage of the CNV and the lack of obvious damage to other structures, it is likely that the condition will progress and lead to further vision loss. However, the rate of progression can vary and it is difficult to predict exactly how quickly the condition will worsen. It is important for the patient to be closely monitored by an ophthalmologist to monitor for any changes in the condition and to determine the best course of treatment.

B

<table>
<tr><td colspan="2"><b>Report</b>
<b>Finding:</b> 1.In the early stage of left eye angiography, a large area of low fluorescence can be seen in the macular area, with a large number of low fluorescence can be seen in the macular area, and the retinal vessels is visible in the late stage, and there is no obvious abnormal fluorescence in the choroidal vessels in the optic disc and retinal vessels. 2.In the late stage of left eye angiography, scattered dot-like drusen can be seen in the peripheral area of the optic disc, and the optic disc and retinal vessels in the left eye.
<b>Impression:</b> drusen, atrophy</td></tr>
<tr><td rowspan="2"><b>Score 3</b></td><td><b>Question:</b> Will the contrast dye used in angiography remain in my body?</td></tr>
<tr><td><b>Answer:</b> Based on the information provided, it is likely that the contrast dye used in angiography will be excreted from your body.
In the early stage of left eye angiography, there is no obvious abnormal fluorescence in the choroidal vessels in the optic disc and retinal vessels, which suggests that the contrast dye has not accumulated in these areas. Additionally, the presence of scattered dot-like drusen in the peripheral area of the optic disc and retinal vessels in the late stage of left eye angiography suggests that the contrast dye has been metabolized and excreted from the body.
It is important to note that the contrast dye used in angiography is typically a non-ionic, water-soluble compound that is designed to be quickly excreted from the body. The dye is injected into a blood vessel in the arm or hand, and then travels through the bloodstream to the retina, where it highlights the blood vessels and allows for clear imaging. Once the imaging is complete, the dye is excreted from the body through the kidneys and liver, and is typically eliminated within 24-48 hours.
However, it is important to follow your healthcare provider's instructions regarding any post-procedure care and follow-up appointments to ensure that the contrast dye has been completely excreted from your body and to monitor for any potential complications.</td></tr>
<tr><td rowspan="2"><b>Score 2</b></td><td><b>Question:</b> What is your suggestion for me?</td></tr>
<tr><td><b>Answer:</b> Based on the information you provided, it seems that you have early stage age-related macular degeneration (AMD) in your left eye, with evidence of drusen and atrophy.
Here are my suggestions for you:
1. Follow-up monitoring: It is essential to monitor your condition regularly with follow-up angiography and optical coherence tomography (OCT) scans to track any changes in your retina and to detect any progression of the disease.
2. Lifestyle modifications: Maintain a healthy lifestyle, including a balanced diet, regular exercise, and avoiding smoking and excessive alcohol consumption.
3. Anti-vascular endothelial growth factor (VEGF) injections: Consider initiating anti-VEGF injections, such as ranibizumab or bevacizumab, to slow down the progression of the disease. These medications can help reduce the number of drusen and slow down the atrophy of the retina.
4. Vitamin supplements: Consider taking vitamin supplements, such as vitamin C, E, and beta-carotene, to help slow down the progression of AMD.
5. Monitor the right eye: Since you have early stage AMD in your left eye, it is essential to monitor your right eye regularly to detect any early signs of the disease.
It is essential to consult with your retina specialist or ophthalmologist to discuss the best course of treatment for your specific condition. They can help you determine the most appropriate treatment based on your individual needs and the severity of your disease.</td></tr>
</table>

C

Report

Finding: 1.There is a large area of choroidal vascular dilation in the macular area, with a large area of low fluorescence in the macular area and increased permeability in the late stage of angiography. 2.There is no obvious abnormalities in the choroidal vessels in the mid-peripheral and peripheral choroidal vessels in the late stage of angiography. 3.No obvious abnormalities were observed in the optic disc and retinal vessels.

Impression: drusen, detachment, Pigment Epithelial Detachment (PED)

Score 1

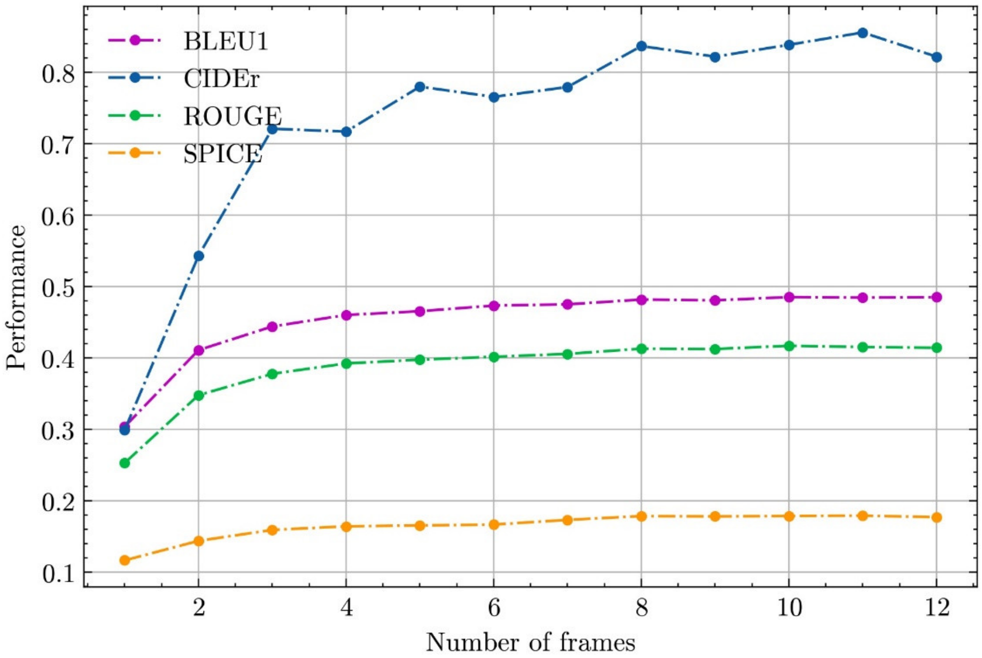
Question: How can I prevent this condition?

Answer: Based on the information provided, it appears that you have a condition called macular edema with choroidal vascular dilation and increased permeability, which is a common complication of diabetic retinopathy. Here are some steps you can take to prevent this condition from worsening and potentially improve your vision:

1. Control your blood sugar levels: As diabetic retinopathy is caused by high blood sugar levels, controlling your blood sugar levels is crucial in preventing the progression of the disease. Monitor your blood sugar levels regularly and work with your healthcare provider to adjust your medication and diet to keep your levels within a healthy range.
2. Follow a healthy diet: Eating a healthy, balanced diet that is low in sugar and refined carbohydrates can help you control your blood sugar levels and prevent the progression of diabetic retinopathy. Include plenty of fruits, vegetables, whole grains, and lean protein sources in your diet.
3. Exercise regularly: Regular exercise can help improve your overall health and reduce your risk of complications from diabetes, including diabetic retinopathy. Aim for at least 30 minutes of moderate-intensity exercise per day.
4. Quit smoking: Smoking is a major risk factor for diabetic retinopathy, so quitting smoking can help reduce your risk of developing the disease.
5. Monitor your blood pressure and cholesterol levels: High blood pressure and high cholesterol levels can contribute to the development of diabetic retinopathy, so monitoring and controlling these factors can help prevent the progression of the disease.
6. Get regular eye exams: Regular eye exams can help detect diabetic retinopathy early, when it is easier to treat and manage. Your healthcare provider may recommend regular eye exams every 6-12 months to monitor your condition.
7. Consider laser surgery: If you have advanced diabetic retinopathy with macular edema and choroidal vascular dilation, your healthcare provider may recommend laser surgery to reduce the risk of vision loss. Laser surgery can help reduce the amount of fluid in the retina and improve vision.

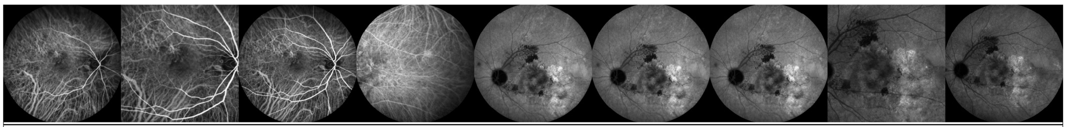
It is important to work closely with your healthcare provider to develop a comprehensive treatment plan that addresses your specific needs and helps prevent the progression of diabetic retinopathy. With proper management and care, it is possible to slow or stop the progression of the disease and preserve your vision.

Supplemental figure S2. Impact of input image number on the report generation performance of ICGA-GPT. ICGA=indocyanine green angiography, GPT=generative pre-trained transformer.



Supplemental figure S3. Demonstration of our model. A. Stage 1: input image, ground truth, and model report generation. B. Stage 2: question answering. User can ask questions in free form, the model answers questions based on the previous generated report and LLM's prior knowledge. Blue Highlight: Corresponds to accurate diagnosis matches. Yellow Highlight: Indicates congruent angiographic and condition descriptions. Red Text: Marks incorrect information in the model's response. Blue Text: Presents supplementary predicted information (not in the manual report).

A



Ground truth

造影所见：1 左眼造影期间于黄斑区可见斑片状脉络膜异常分支血管网，中后期其边缘可见囊样强荧光，造影晚期未见明显渗漏；2 左眼造影期间于病灶周围可见片状渗出性遮蔽荧光；3 左眼造影期间中周部点状玻璃膜疣性强荧光。

造影提示：1 左眼黄斑区BVN伴静止性息肉灶；2 左眼中周部玻璃膜疣染色。

Chinese prediction

造影所见：1 左眼造影早期于黄斑区可见片状脉络膜异常分支血管网及其末梢囊样扩张的息肉状强荧光，造影晚期呈冲刷现象；2 左眼造影期间于黄斑区可见一约3PD边界清楚的不规则弱荧光，随造影时间延长，上半部荧光渐强，下半部呈遮蔽荧光，造影晚期形态大小不变；3 左眼造影中晚期可见后极部及各中周部玻璃膜疣性强荧光。

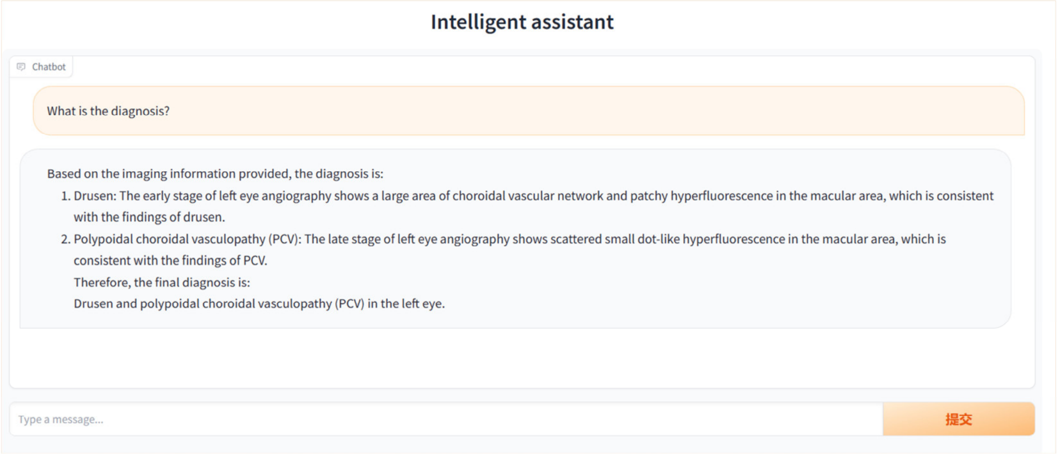
造影提示：1 左眼黄斑区BVN伴息肉灶；2 左眼黄斑区浆血性PED；3 左眼玻璃膜疣染色

English prediction

Finding: 1. In the early stage of left eye angiography, a large area of choroidal vascular network and patchy hypofluorescence can be seen in the macular area, with a patchy hyperfluorescence can be seen in the late stage, and dye leakage of the angiography time, and in the late stage, there is no obvious abnormal fluorescence in the macular area. 2. In the late stage of left eye angiography, scattered small dot-like hyperfluorescence can be seen in the macular area, and the late stage, there is no obvious abnormal fluorescence in the optic disc and retinal vessels.

Impression: drusen OS, polypoidal choroidal vasculopathy OS

B



What is the diagnosis?

Based on the imaging information provided, the diagnosis is:

1. Drusen: The early stage of left eye angiography shows a large area of choroidal vascular network and patchy hyperfluorescence in the macular area, which is consistent with the findings of drusen.
2. Polypoidal choroidal vasculopathy (PCV): The late stage of left eye angiography shows scattered small dot-like hyperfluorescence in the macular area, which is consistent with the findings of PCV.
   Therefore, the final diagnosis is:
   Drusen and polypoidal choroidal vasculopathy (PCV) in the left eye.

Provide insights based on the following components for *RAE submission requirement (no more than 100 words):

Originality:
This study successfully developed and evaluated an innovative model related to indocyanine green angiography (ICGA) image interpretation by combining a multimodality transformer architecture and a large language model.

Significance:
This model not only provides ICGA report generation but also enables subsequent question-answering, demonstrating a promising adjunct to enhance medical image interpretation for the benefit of healthcare professionals and patients.

Rigour:
The generated ICGA reports and answers were under qualitative evaluation using language-based metrics and disease-based metrics. Additionally, three experienced ophthalmologists conducted subjective evaluations using a 5-point scale.