nature communications

Article

https://doi.org/10.1038/s41467-025-58801-7

A concept-based interpretable model for the diagnosis of choroid neoplasias using multimodal data

Received: 26 August 2024

Accepted: 2 April 2025

Published online: 13 April 2025

Check for updates

Yifan Wu[1,9], Yang Liu[2,9], Yue Yang[1], Michael S. Yao[1], Wenli Yang[3,4,5], Xuehui Shi[3,4,5], Lihong Yang[3,4,5], Dongjun Li[3,4,5], Yueming Liu[3,4,5], Shiyi Yin[3,4,5], Chunyan Lei[6], Meixia Zhang[6], James C. Gee[1], Xuan Yang[3,4,5], Wenbin Wei[3,4,5] & Shi Gu[2,7,8]

Diagnosing rare diseases remains a critical challenge in clinical practice, often requiring specialist expertise. Despite the promising potential of machine learning, the scarcity of data on rare diseases and the need for interpretable, reliable artificial intelligence (AI) models complicates development. This study introduces a multimodal concept-based interpretable model tailored to distinguish uveal melanoma (0.4-0.6 per million in Asians) from hemangioma and metastatic carcinoma following the clinical practice. We collected a comprehensive dataset on Asians to date on choroid neoplasm imaging with radiological reports, encompassing over 750 patients from 2013 to 2019. Our model integrates domain expert insights from radiological reports and differentiates between three types of choroidal tumors, achieving an F_1 score of 0.91. This performance not only matches senior ophthalmologists but also improves the diagnostic accuracy of less experienced clinicians by 42%. The results underscore the potential of interpretable AI to enhance rare disease diagnosis and pave the way for future advancements in medical AI.

Recent advancements in machine learning and deep neural networks have accelerated the development of computer-aided diagnostic (CAD) methods over the past decade[1]. For common diseases amenable to automated diagnoses with large publicly available datasets, deep learning-based models are able to perform comparably to radiologists across a variety of diagnostic tasks[2–6]. However, for rare diseases, particularly rare oncologic diseases, the development of CAD methods remains underexplored. Although individually uncommon, rare cancers collectively account for ~24% of all new cancer cases[7] and present substantial challenges due to the lack of corresponding widespread clinical expertise in managing these conditions. Thus, improving the quality of artificial intelligence (AI)-based solutions for rare cancer diseases will contribute significantly to public health[8–10].

The development of AI tools for rare diseases is impeded by two significant challenges compared to those for common diseases. Firstly, the scarcity of high-quality datasets impedes the advancement of learning-based approaches. This issue stems from the low prevalence of these diseases, the prohibitive cost of professional annotations, and potential incompatibilities between clinical and research protocols. Such constraints associated with the clinical management of rare

[1]University of Pennsylvania, Philadelphia, PA, USA. [2]University of Electronic Science and Technology of China, Chengdu, China. [3]Beijing Tongren Eye Center, Beijing Tongren Hospital, Capital Medical University, Beijing, China. [4]Beijing Key Laboratory of Intraocular Tumor Diagnosis and Treatment, Beijing Tongren Hospital, Capital Medical University, Beijing, China. [5]Beijing Ophthalmology and Visual Sciences Key Lab, Beijing Tongren Hospital, Capital Medical University, Beijing, China. [6]Department of Ophthalmology and Research Laboratory of Macular Disease, West China Hospital, Sichuan University, Chengdu, China. [7]College of Computer Science and Technology, Zhejiang University, Hangzhou, China. [8]State Key Laboratory of Brain Machine Intelligence, Zhejiang University, Hangzhou, China. [9]These authors contributed equally: Yifan Wu, Yang Liu. e-mail: yangxuan153@126.com; weiwenbintr@163.com; gus@uestc.edu.cn

diseases make it difficult to leverage data-based approaches[11,12], such as Contrastive Language-Image Pre-training (CLIP)[13] and generative Large Language Models (LLMs). Secondly, the interpretability of the model is crucial for making these solutions trustworthy and for their integration into clinical practice[14–16]. General practitioners individually manage only a small number of cases of any particular rare disease given their low prevalence, making it essential for analytical tools to not only provide accurate predictions but also produce interpretable explanations in alignment with the insights of experienced specialists to support comprehensive clinical management[17]. Interpretability in this context encompasses two crucial dimensions: (1) understanding how the model arrives at a particular prediction; and (2) ensuring that this process is presented in a human-readable format[18,19].

Here we explore our methodology for designing machine learning models specifically tailored for the diagnosis of rare diseases, with a particular focus on uveal melanoma. This rare cancer originates from various components of the uveal tract in the eye, such as the iris, ciliary body, and choroid[20,21]. The incidence of this disease is 6 per million in European populations[22] and only 0.6 per million in Asia[23]. Furthermore, the prognosis for uveal melanoma patients is often poor, largely because of the high risk of metastasis at the time of diagnosis[24]. As a result, these melanomas frequently go undetected in routine clinical evaluations. A crucial initial step is distinguishing choroidal melanoma (i.e., the most prevalent subtype of uveal melanoma) from other similar conditions like metastatic carcinoma and hemangioma, which occur in the choroid of the fundus and typically present as solitary tumors. These conditions can exhibit similar clinical symptoms and overlapping imaging features in early presentation[25]. In this work, we aim to build an interpretable computer-aided system to differentiate between choroidal melanoma, metastatic carcinoma, and hemangioma.

Considering the tumor location and the importance of the eyes, the diagnosis of choroid neoplasias critically depends on imaging techniques rather than the biopsy for other tumors[20,26]. Initial diagnosis requires a detailed fundoscopic examination with an expert clinician followed by additional imaging techniques such as ocular ultrasound (US), fluorescein angiography (FA), and indocyanine green angiography (ICGA), as well as magnetic resonance imaging (MRI) for confirmation and prognostication[27–30]. While these imaging techniques are accessible in many general eye hospitals, domain-specific radiologists and expert ophthalmologists specializing in managing choroidal neoplasias are few and far between, further complicating diagnostic workup[29]. Given the poor prognosis associated with choroidal melanomas and the consequent need for timely diagnosis and treatment, it is crucial to have high confidence in a diagnosis of choroidal neoplasias prior to definitive intervention.

In this work, we propose an interpretable model that encodes the expertise of specialized clinicians into AI to produce human-understandable diagnostic outputs (Fig. 1). To establish such a pipeline for the automated interpretable diagnosis of choroidal neoplasias, we need to address key challenges in data curation and model verification. First, to enable our work, we collect and release a carefully

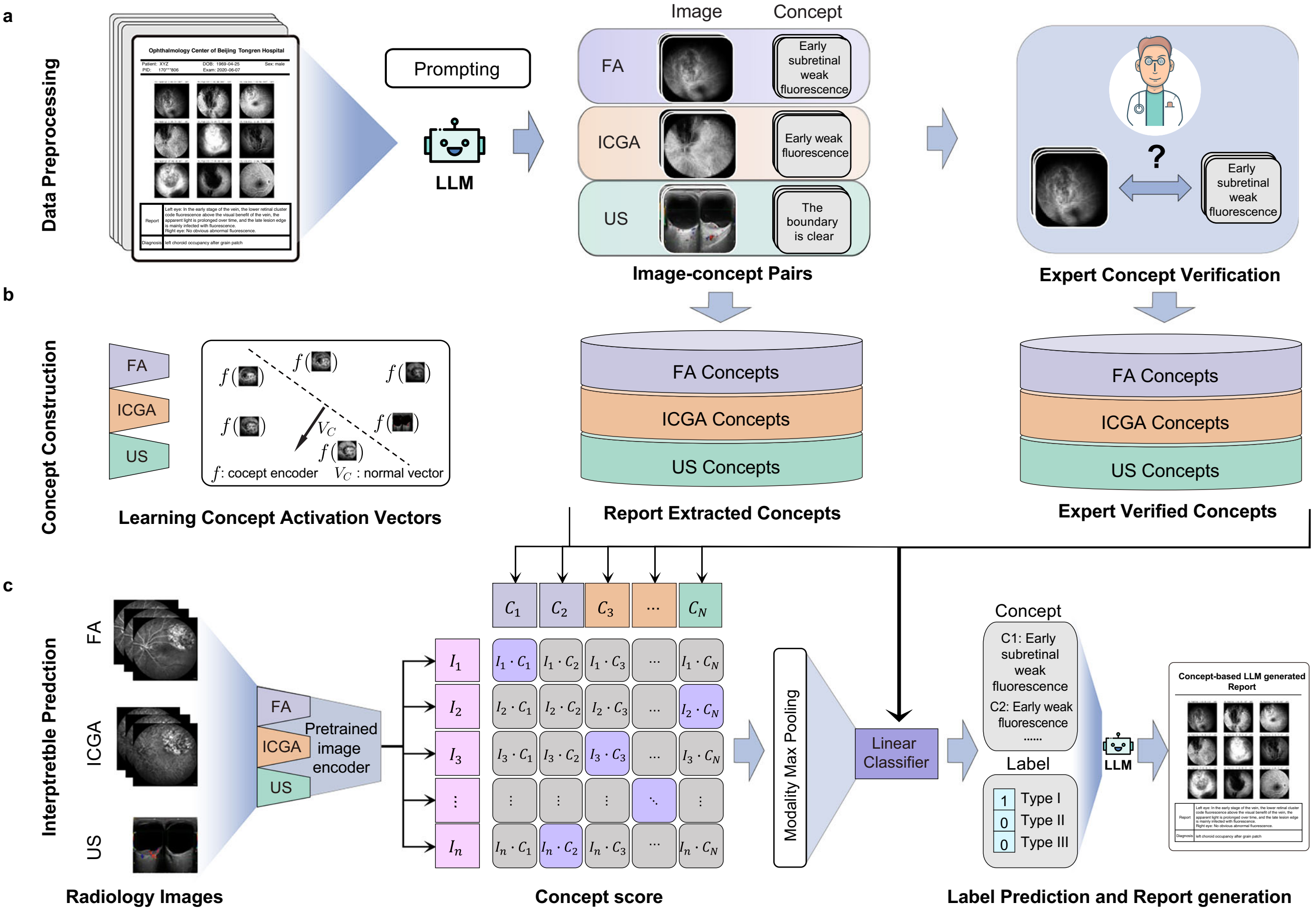
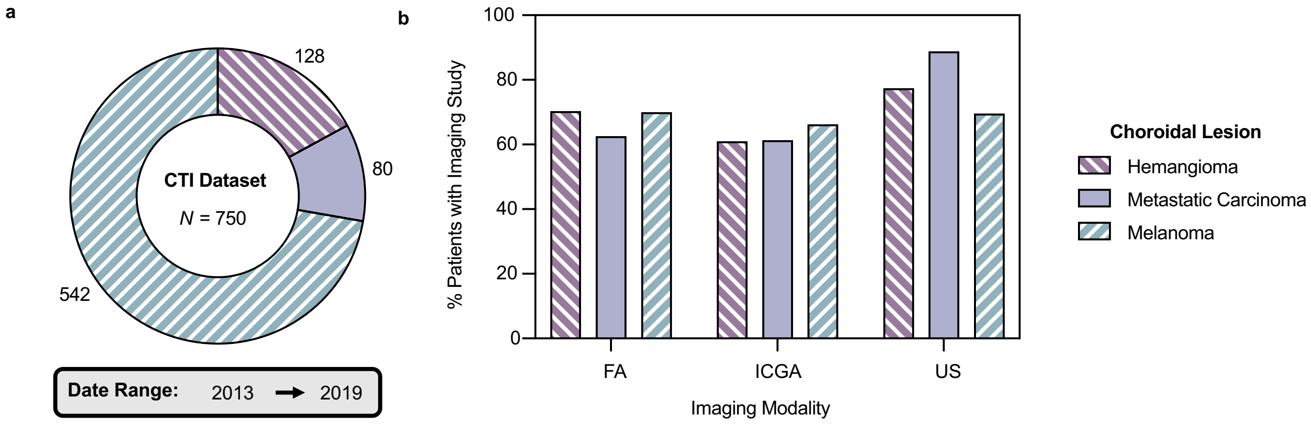

a
Data Preprocessing
Ophthalmology Center of Beijing Tongren Hospital
Prompting
LLM
Image
Concept
FA
ICGA
US
Early subretinal weak fluorescence
Early weak fluorescence
The boundary is clear
Image-concept Pairs
?
Early subretinal weak fluorescence
Expert Concept Verification
b
Concept Construction
FA
ICGA
US
f
V_C
f: cocept encoder
V_C : normal vector
Learning Concept Activation Vectors
FA Concepts
ICGA Concepts
US Concepts
Report Extracted Concepts
FA Concepts
ICGA Concepts
US Concepts
Expert Verified Concepts
c
Interptretble Predction
FA
ICGA
US
Radiology Images
FA
ICGA
US
Pretrained image encoder
C_1 C_2 C_3 ⋯ C_N
I_1 I_2 I_3 ⋮ I_n
$I_1 \cdot C_1$ $I_1 \cdot C_2$ $I_1 \cdot C_3$ ⋯ $I_1 \cdot C_N$
$I_n \cdot C_1$ $I_n \cdot C_2$ $I_n \cdot C_3$ ⋯ $I_n \cdot C_N$
Concept score
Modality Max Pooling
Linear Classifier
Concept
C1: Early subretinal weak fluorescence
C2: Early weak fluorescence
......
Label
1 Type I
0 Type II
0 Type III
LLM
Concept-based LLM generated Report
Label Prediction and Report generation

Fig. 1 | Overview of the MMCBM workflow. a Utilizing a large language model (LLM), concept banks are formulated by extracting image-concept pairs from comprehensive medical reports. Senior experts help examine the faithfulness of the image-concept pairs and make corresponding modifications. **b** Based on such pairs, we construct the concept bank by learning concept activation vectors. **c** The model's output stage takes a series of images spanning 1 to 3 modalities. A pretrained image encoder is employed to convert these images into tokenized features. Subsequent calculations produce concept scores. The model then delivers an explainable prediction, spotlighting the diagnostic evidence. Moreover, it crafts an interpretative report, enhancing the transparency of the diagnostic process.



c

	Training Dataset				Test Dataset
Choroidal Lesion	FA Only	ICGA Only	US Only	Multi-Modal	Multi-Modal
Hemangioma	37	25	46	42	11
Metastatic Carcinoma	12	12	37	30	8
Melanoma	185	165	183	155	39

Fig. 2 | Statistics of the CTI[49] dataset. a The CTI dataset is composed of 750 patients: 542 with melanoma, 128 with hemangioma, and 80 with metastatic carcinoma, collected from 2013 to 2019. **b** Proportions of patients with hemangioma, metastatic carcinoma, and melanoma imaged by fluorescein angiography (FA), indocyanine green angiography (ICGA), and ultrasound (US). Source data are provided as a Source Data file. **c** Split of imaging studies in the training and test datasets across various imaging modalities: 20% of the multimodal data (MM), representing patients imaged with all three modalities, is set aside for testing. The remaining 80% of MM and all non-MM data were allocated for training using five-fold cross-validation.

curated multimodal dataset of uveal oncologic images to train classifier models that accurately differentiate choroidal melanomas from other clinically similar diseases. To our knowledge, this dataset represents a comprehensive collection of choroidal melanomas in Asian populations. We then use this dataset to develop the Multimodal Medical Concept Bottleneck Model (MMCBM), a domain knowledge-enhanced model that predicts interpretable classifications from patient data. MMCBM supports a human-in-the-loop mechanism to learn from the feedback provided by domain experts. We find that MMCBM not only provides accurate classifications but also offers interpretable visual feature concepts on primary imaging modalities that explain their reasoning process. We show that these concepts align well with senior doctors and provide substantial assistance for trainees and less experienced clinicians to diagnose choroidal neoplasias more accurately. Our methodology leverages the extensive knowledge in clinical reports to offer a pathway towards building interpretable models for diagnosing rare diseases.

Results

Dataset description

To support the development of interpretable models for diagnosing choroidal tumors, we built the Choroid Tri-Modal Imaging (CTI) dataset, an anonymized, multimodal, and annotated collection of medical images from Beijing Tongren Hospital (2013–2019) encompassing fluorescence angiography (FA), indocyanine green angiography (ICGA), and ocular ultrasound (US) images. The construction of this dataset was approved by the Ethics Committee of Beijing Tongren Hospital. The CTI dataset includes images from patients diagnosed with benign hemangioma, secondary metastatic carcinoma to the eye, or primary choroidal melanoma (Fig. 2 and Supplementary Fig. 9c), comprising a total of 750 subjects, with 344 female subjects having an average age of 47.3 ± 13.0 years and 406 male subjects having an average age of 47.0 ± 13.5 years (see Supplementary Fig. 1 and Supplementary Tables 1, 2 for the detailed distribution). Specific for each category, there are 542 patients with choroidal melanoma (FA: 379, ICGA: 359, US: 377), 128 patients with choroidal hemangioma (FA: 90, ICGA: 78, US: 99), and 80 patients with choroidal metastatic carcinoma (FA: 50, ICGA: 49, US: 71). The numbers indicate the quantity of imaging studies for each specific imaging modality. Note that not every patient has images across all modalities. We refer to the subset where patients have all three modalities as MultiModal (MM) data and reserve 20% of this MM data as a hold-out test set. In the MM data training split, 97 patients have anonymized reports for all three modalities, describing the radiological features observed in the images.

Baseline black-box model

We first seek to build baseline black-box machine learning models. Inspired by recent success in natural image processing[31] and the medical application with multimodal data[32], our baseline black-box model comprises three separate modality-specific encoders trained to encode corresponding imaging study inputs into intermediate lower-dimensional representations. The encoder output (or outputs, if multiple imaging studies of different modalities are available for a given patient) is then passed to an attention-pooling block[33] and a subsequent dense layer to yield the final classification prediction (Supplementary Fig. 9a). We refer to this model architecture as the pretrained multimodal classifier. Our baseline model performs accurately across different input image modalities, validating the feasibility of deep-learning models for this clinical problem. Using FA imaging studies alone, the pretrained classifier achieves an F_1 score of 78.3% (95% CI: 74.0–81.7%); using ICGA studies alone, it achieves an F_1 score of 85.9% (95% CI: 83.7–88.2%); and using US studies alone, it achieves an F_1 score of 72.1% (95% CI: 67.1–76.7%). When using all three imaging studies together, the baseline classifier attains an F_1 score of 89.2% (95% CI: 87.9–90.6%). Additional results are included in Supplementary Table 4. Our results show that using multimodal inputs leads to more accurate models than those leveraging any individual imaging study as input alone. However, while the pretrained multimodal classifier demonstrates impressive performance, it is impossible to interrogate the model's predictions for human clinicians to interpret—a key limitation of existing black-box approaches.

Trustworthy interpretable framework: MMCBM

The lack of interpretability in the baseline pretrained classification model is a common limitation of many modern AI tools. To address the need for trustworthiness in medical diagnostics, we sought to engineer a framework with interpretability integrated directly as a part of the model design. Our approach, the Multimodal Medical Concept Bottleneck Model (MMCBM), is designed to integrate human expertise directly into the diagnostic process, especially crucial in high-stakes medical contexts. The core idea is to utilize the seasoned knowledge of ophthalmology experts to refine how our model processes and interprets image data. We achieve this by aligning the model's interpretation of images with the kinds of visual patterns and diagnostic criteria that ophthalmologists routinely use to make their diagnoses. This alignment allows the model's outputs to be easily understandable: they are presented as combinations of recognizable visual elements and clinical findings. These are the same elements that clinicians use as evidence in their diagnostic decisions and are also employed as teaching tools in training scenarios. This ensures that our model not only aids in diagnosis but does so in a way that is transparent and educative for medical professionals. These representations are referred to as *concepts*.

Concept construction and grounding

Using medical reports as the knowledge database, we prompt GPT-4[34] to extract concepts from reports and construct a bank of concepts containing phrases related to imaging findings of choroidal tumors. For instance, a description in a fluorescein angiography (FA) report states, "In the venous phase, a clustered hypofluorescence under the subretinal can be seen in the temporal part of the macula. Fluorescence increases with time, and lesions are dominated by fluorescent staining at the late stage." The extracted concepts for this FA study include "Clustered Hypofluorescence During Venous Phase", "Globally Increasing Fluorescence Intensity", and "Late-Stage Staining". After extracting concepts from the reports of 97 patients, we use GPT-4 to aggregate semantically similar concepts, ensuring each concept's uniqueness and relevance. The final concept bank consists of 47 concepts for FA, 30 for ICGA, and 26 for US, with an average of 3 concepts for FA, 2 for ICGA, and 5 for US per patient. The comprehensive list of all $N = 103$ concepts is presented in Supplementary Table 6. To validate that the concepts extracted by the LLM accurately represented real-world clinical reasoning, two senior ophthalmologists specializing in diagnosing and managing choroidal tumors at Beijing Tongren Hospital were asked to verify and amend the concepts. The changes made by the experts are shown in Supplementary Fig. 10. Quantitatively, the initial concept bank constructed by GPT-4 was assessed to be reasonable and relevant, requiring only minor modifications: 5 concepts were removed, 8 new ones were added to the FA category, 4 to the ICGA category, and no changes to the US category.

To ground concepts as feature embeddings, we employed support vector machines (SVMs) for concept-level binary classification. We used image representations from a pretrained model as input and binary vectors derived from the concept construction process as labels. Images associated with assigned concepts were used as positive samples, and all other images were used as negative samples. The classification hyperplane vector from each SVM serves as the concept's representation, which we refer to as concept activation vectors (CAVs)[35]. Subsequently, in MMCBM, an image is projected into the space of concepts to estimate the input image alignment with any given modality-specific concept. The alignment scores are then input into a linear classifier to predict the relative probabilities of each of the three targeted choroidal diseases. Figure 3a shows this process and shows the top-*k* concepts derived from concept scores to explain the model's predictions.

Comparison of model performance between MMCBM and black-box model

A common critique of interpretable machine learning models is that enforcing priors on the model, such as requiring input images to align with concept activation vectors, is equivalent to adding additional regularization to the hypothesis space[18]. Such constraints may adversely impact the performance of trained models[36,37]. To this end, we seek to evaluate the classification performance of our MMCBM model against the black-box pretrained multimodal classifier baseline (Fig. 3). On the MM test set, MMCBM achieves an overall classification F_1 score of 91.0% (95% CI: 88.2– 93.4%), and the performance of the baseline black-box model is (89.2%; 95% CI: 87.9–90.6%). The two-sample *t*-tests suggest that the performance between MMCBM and the black-box model are not statistically different. Additionally, comparing classifier performance across unimodal imaging inputs reveals no statistically significant differences in classification metrics (Supplementary Table 4 and Supplementary Fig. 2). This indicates that our MMCBM framework matches the performance of black-box approaches in automating the diagnosis of rare choroidal tumors according to clinically relevant metrics. Additionally, ablation studies were conducted: Supplementary Fig. 2 for encoder size, Supplementary Fig. 4 for the number of reports, Supplementary Figs. 3, 5 for the number of concepts and Supplementary Fig. 6 for few-shot learning.

Evaluation of the generalizability of MMCBM

To assess the generalizability and robustness of our models trained on the CTI dataset, we re-trained the MMCBM and black-box models on all the 750 subjects and conducted validations on two additional datasets: an internal independent dataset from Beijing Tongren Hospital collected between 2020 and 2023 (i.e., disjoint in time from the CTI dataset), and an external test set from West China Hospital collected between 2023 and 2024 (i.e., disjoint in both time and patient population from the CTI dataset), are shown in Supplementary Table 3. Each dataset is similarly structured in three modalities and focuses on the same types of choroidal tumors, ensuring compatibility and relevance for our validation efforts. The internal independent dataset comprises 83 cases, including 26 cases of choroidal hemangioma, 44 cases of primary choroidal melanoma, and 13 cases of secondary metastatic carcinoma. The external test set includes 43 cases, with a balanced distribution of 16 cases each of hemangioma and melanoma, and 11 cases of metastatic carcinoma. For the new dataset from Tongren Hospital, the MMCBM model achieves an accuracy of 92.8%, a sensitivity of 90.3%, and an F1-score of 90.3%. The black-box (BB) model attains a comparable accuracy of 92.8%, a slightly lower sensitivity of 87.2% (BB vs. MMCBM: $z = 5.70$, $p < 1.0 \times e^{-5}$), and an F1-score of 89.6% (BB vs. MMCBM: $z = 1.30$, $p = 0.19$). For the test dataset from West China Hospital, the MMCBM model achieves an accuracy of 88.6%, an sensitivity of 87.5%, and an F1-score of 87.9%—outperforming the black-box model with a lower accuracy of 77.2% (BB vs. MMCBM: $z = 9.20$, $p < 1.0 \times e^{-5}$), a lower sensitivity of 73.6% (BB vs. MMCBM: $z = 10.72$, $p < 1.0 \times e^{-5}$) and a lower F1-score of 72.4% (BB vs. MMCBM: $z = 11.90$, $p < 1.0 \times e^{-5}$). Of note, our MMCBM model has only a 4% drop in accuracy on the West China Hospital dataset, which was less than the 15% drop in accuracy for the black-box model. Our results support the generalizability of the MMCBM model when compared to the black-box models, especially when there is a distribution shift between model training and test populations.

Test-time random intervention with Oracle

The ability to intervene in MMCBM enables richer and more interactive engagement for human users. For example, if a clinician

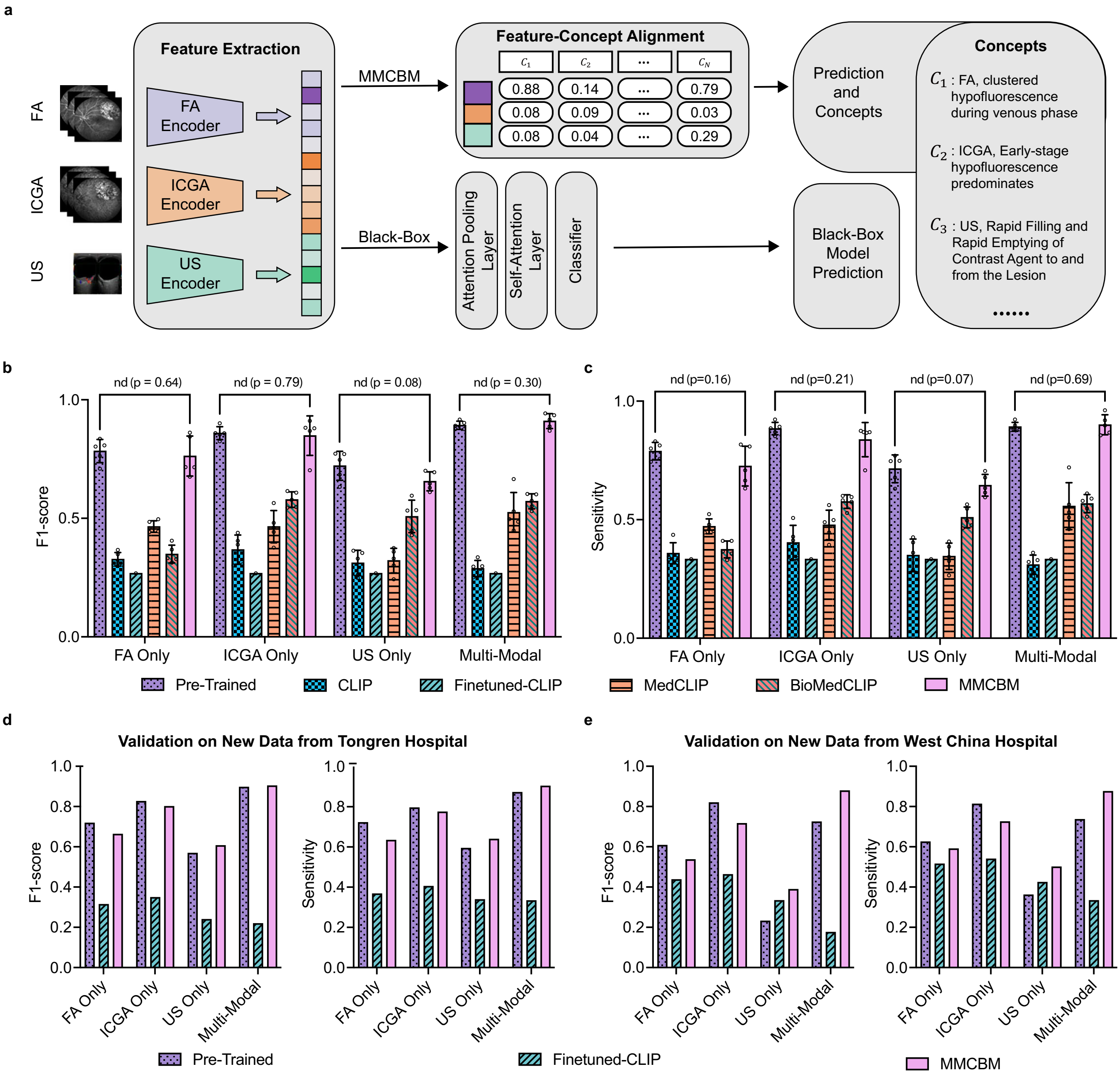
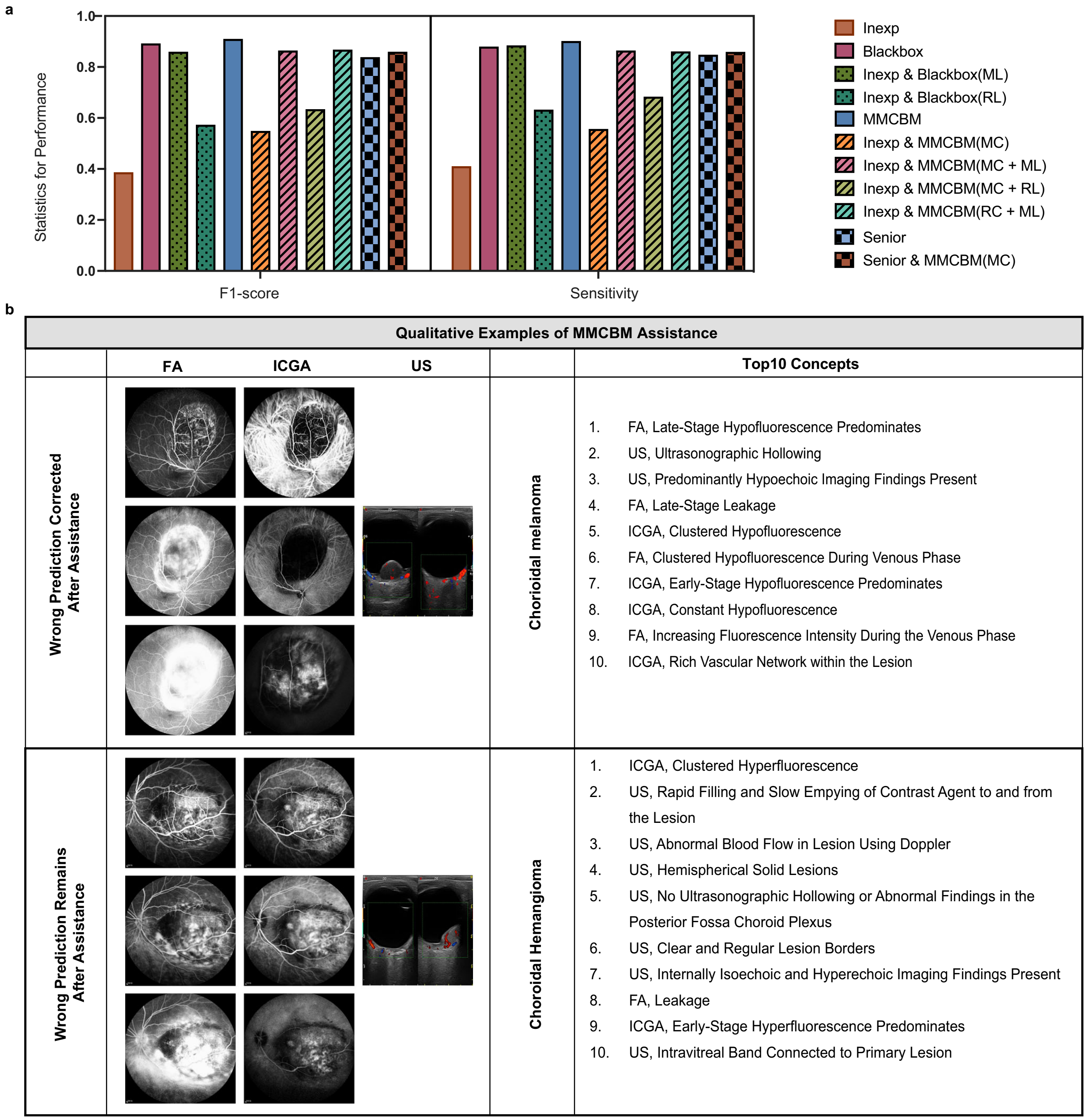


Fig. 3 | Multimodal medical concept bottleneck model (MMCBM). Black-box models, such as the pretrained classifier, learn directly from the encoded image features and output a single model prediction without any insight as to how the prediction was computed. In contrast, the MMCBM shown in (**a**) instead represents encoded features by their alignment with key medical concepts derived from domain experts. This allows MMCBM to return not only its prediction but also the top-*k* activated concepts that best describe the input images, giving insight into how the model arrived at its prediction. Comparing both the classification (**b**) F1-score and **c** sensitivity of the models, there is no statistically significant difference between black-box models and MMCBM across all sets of imaging inputs. MMCBM concepts also outperform features derived from CLIP-based models, highlighting the importance of sourcing prior knowledge from domain experts. The data were presented as mean ± SD, with error bars indicating the standard deviation from $n = 5$ independent replicates. Statistical significance is analyzed using an unpaired two-sided *t*-test. To demonstrate the generalization ability of the MMCBM, we examine the model performance on two additional datasets. One (**d**) presents the classification performance of models using data from Tongren Hospital, albeit from outside the original collection period. Meanwhile. The other (**e**) depicts the results of the external dataset from West China Hospital of Sichuan University, where variations in both scan protocols and reporting styles are evident. Source data are provided as a Source Data file.

disagrees with the model's prediction, they can inspect the predicted concepts, correct erroneous concept scores, and analyze how the model's output changes in response to these adjustments. In clinical settings, domain experts interacting with the model may intervene to rectify potentially inaccurate concept values predicted by the system. To study this setting, we use an oracle that can query the existence of any concept for a test input. In Fig. 4b, we show examples of interventions that lead to the corrected prediction. Consider a patient with three-modality images, where the multimodal concepts are provided by the MMCBM. For each modality, we obtain the concepts present in the images through experts' annotations. We classify these annotated concepts as positive concepts and the remaining concepts as negative. To quantify the presence or absence of a concept, we analyze the distribution of concept scores from the training data. We designate the top 5% score as indicating presence (active) and the bottom 5% score as indicating absence (inactive). This approach allows us to intervene on concepts predicted by the MMCBM as active if they are present in the experts' annotations. However, since the experts' annotations do not include importance rankings, we randomly



Qualitative Examples of MMCBM Assistance

	FA	ICGA	US		Top10 Concepts
Wrong Prediction Corrected After Assistance				Choroidal melanoma	1. FA, Late-Stage Hypofluorescence Predominates 2. US, Ultrasonographic Hollowing 3. US, Predominantly Hypoechoic Imaging Findings Present 4. FA, Late-Stage Leakage 5. ICGA, Clustered Hypofluorescence 6. FA, Clustered Hypofluorescence During Venous Phase 7. ICGA, Early-Stage Hypofluorescence Predominates 8. ICGA, Constant Hypofluorescence 9. FA, Increasing Fluorescence Intensity During the Venous Phase 10. ICGA, Rich Vascular Network within the Lesion
Wrong Prediction Remains After Assistance				Choroidal Hemangioma	1. ICGA, Clustered Hyperfluorescence 2. US, Rapid Filling and Slow Empying of Contrast Agent to and from the Lesion 3. US, Abnormal Blood Flow in Lesion Using Doppler 4. US, Hemispherical Solid Lesions 5. US, No Ultrasonographic Hollowing or Abnormal Findings in the Posterior Fossa Choroid Plexus 6. US, Clear and Regular Lesion Borders 7. US, Internally Isoechoic and Hyperechoic Imaging Findings Present 8. FA, Leakage 9. ICGA, Early-Stage Hyperfluorescence Predominates 10. US, Intravitreal Band Connected to Primary Lesion

Fig. 4 | Test-time random intervention with Oracle. a Comparison of evaluation performance as the number of random interventions increases. For each intervention, concepts are randomly selected from the experts' annotated concepts in the test set, covering multiple modalities. The purple dot indicates the mean performance, with the error bar showing the standard deviation. The small black dots represent individual samples, each reflecting different compositions of positive and negative concepts. Source data are provided as a Source Data file. **b** Examples of successful random interventions illustrating the use of positive or negative concepts, as randomly sampled from experts' concept annotations, to adjust the model's concept scores. These adjustments effectively correct the model's initial predictions, demonstrating how targeted concept interventions can improve overall performance.

sample k_{pos} positive and k_{neg} negative concepts. We then adjust their corresponding concept scores to active or inactive statuses to observe the impact on classification performance. For each patient, we repeat this intervention process 50 times and record the best one as an oracle. In Fig. 4a, we present the results of test-time oracle intervention. Notably, both sensitivity and F_1-score surpass those of the non-intervened model when fewer than eight concepts per modality are intervened, with performance peaking at 3–4 concepts. This suggests that moderate concept intervention enhances the model's understanding of the task, improving performance, while excessive intervention introduces noise and disrupts learned representations, reducing effectiveness. In high-stakes environments such as medicine, this capability empowers experts to interactively guide the model, ensuring that critical concepts are accurately represented and reducing the likelihood of erroneous predictions. The detailed results for various combinations of positive and negative concepts are provided in Supplementary Table 5.

Integration of MMCBM in clinical workflows

We have shown that the MMCBM effectively leverages prior knowledge from domain experts to represent input data aligned with interpretable concepts. However, it remains unknown whether our framework can provide real-world utility in augmenting existing clinical workflows. To investigate the applications of MMCBM in clinical practice, we recruited the help of 8 doctors from Beijing Tongren Hospital: two senior ophthalmologists specializing in the diagnosis and management of choroidal melanomas, and six resident ophthalmologists in training. We assess the diagnostic performance of ophthalmologists alone against ophthalmologists with our trained MMCBM model. In order to avoid the memorization of the seen cases, the time lapse between the experiments w./w.o., the generated concepts is 2 months. The ophthalmologists leveraging our MMCBM model for diagnostic workflow augmentation had access to the top 10 activated concepts from the MMCBM concept bank and were able to adjust the confidence scores of the concepts based on their judgment. This human-in-the-loop interactive feature improves the practical utility of MMCBM in clinical decision-making, fostering a more collaborative and accurate diagnostic process. For the 6 less experienced ophthalmologists, the average accuracy is 51.9%, precision 40.5%, recall 40.9%, and F_1 score 38.5%; with the aid of our MMCBM model, their accuracy improves to 65.5%, precision to 54.3%, recall to 55.5%, and F_1 score to 54.7% (Fig. 5a). The 2 senior ophthalmologists demonstrate a high diagnostic accuracy at baseline of 91.4%, precision of 85.8%, recall of 83.6%, and F_1 score of 84.6%. When augmented with the model's predictions, their performance remains relatively unchanged, with an accuracy of 91.4%, precision of 86%, recall of 85.8%, and F_1 score of 85.7%. In particular, the use of MMCBM improves less experienced doctors' performance by 42% on the F_1 score. Detailed results on separate groups are shown in Supplementary Fig. 11. These results not only validate the quality and precision of the predicted concepts of our MMCBM model but also highlight our model's ability to serve educational purposes by improving the diagnostic accuracy of less experienced doctors for complex and rare diseases.

Comparison of assisted performance via MMCBM and black-box model

To further evaluate the effect of assistance with MMCBM and examine how it is distinct from the black-box model, we designed human verification experiments where six inexperienced doctors were given different assistant results, including attention map, concepts, and prediction probabilities, and were asked to predict the labels. Moreover, in order to evaluate the effect of different ways of assistance more comprehensively, we add a random group with the same test samples, yet the probability or concepts are given randomly. Both random and unrandom groups have 11 subjects for hemangioma, eight subjects for metastatic carcinoma, and 39 subjects for melanoma. In Table 1 that corresponds to the exact numbers in Fig. 5a, we can see that the group with both labels and concepts generated from the model have the best performance. Another interesting observation is that inexperienced doctors are highly likely to be affected by the given label. The final performance highly depends on whether reliable labels are given. However, when the model-generated concepts are given, compared to the group with only randomized labels, the accuracy increased from 62.7 to 70.2% (MC vs. MMCBM: $z = 6.40$, $p < 1.0 \times 10^{-5}$), precision increased from 56.1 to 61.1% (BB vs. MMCBM: $z = 4.20$, $p < 1.0 \times 10^{-5}$), recall increased from 63% to 68.2% (BB vs. MMCBM: $z = 4.38$, $p < 1.0 \times 10^{-5}$), and the F1-score increased from 57.1% to 63.2% (BB vs. MMCBM: $z = 5.04$, $p < 1.0 \times 10^{-5}$).

To evaluate how the assistance works for cases with different levels of challenges for human beings, we calculate the percentage of cases in each category based on whether the evaluator diagnoses it correctly or not before and after different types of assistance. In Table 2, we can first see that the assistance of the black-box model and MMCBM result in similar performance as long as the model-generated labels are given (Row 1 & 2 in Table 2). When the labels are randomly given in the assistance stage, it would highly distort the judgment of the evaluator (Row 3 & 4 in Table 2). Providing concepts to the evaluator may slightly relieve this issue where the possibilities of correcting wrong answers and keeping the true answers both increases. Specially, for the cases where the evaluators made mistakes in the initial stage, they can correct 67.9% with the assistance of concepts and random labels (Row 4 in Table 2), while the rate for black-box models with random labels is 53.6% (Row 3 in Table 2). Another finding is that when the attention map is given with the black-box model, even if the labels are random, the evaluator is more likely to make the correct prediction in the initial stage, increasing from 51.7% (Row 4 in Table 2) to 63.8% (Row 5 in Table 2). At the same time, the possibility of correcting wrong answers decreases from 67.9% (Row 4 in Table 2) to 61.9% (Row 5 in Table 2). When both the model-generated concepts and labels are given, the evaluators are then much more less likely to be distorted for the cases that they made the correct decision in the beginning, where the possibility of keeping right answers increases from 83.8% (Row 5 in Table 2) to 97.3% (Row 6 in Table 2). Overall, we can conclude that the integration of interpretable concepts in the computer aid diagnosis can help relieve the distortion and overreliance of labels from the black-box models with attention maps.

Comparison between MMCBM and alternative feature embedding methods

Given the recent progress made in cross-modality foundation models, it may be possible to leverage existing feature embedding models trained on extensive corpora of medical information to represent input ocular imaging data and concepts. This approach might offer greater generalizability and require less effort than our MMCBM setup. To evaluate this alternative framework, we compared our concept embedding procedure and image feature extraction with those using contrastive language-image pre-training (CLIP)[13] and its biomedical variants, including MedCLIP[38] and BioMedCLIP[39], which are specifically fine-tuned for medical data. Briefly, MedCLIP was fine-tuned on multiple Chest X-ray datasets, while BioMedCLIP underwent fine-tuning on 15 million figure-caption pairs extracted from biomedical research articles in PubMed Central. Our results suggest that all assessed CLIP-based frameworks perform significantly worse than our CAV-based feature extraction method used in our MMCBM framework (Fig. 3b, c). As expected, methods fine-tuned on specialized medical datasets—such as MedCLIP and BioMedCLIP—outperform the generic CLIP model as feature extractors for choroidal disease diagnosis using both multimodal and unimodal image inputs (Fig. 3b, c, MedCLIP: 52.5% (95% CI: 47.2–59.6%), BioMedCLIP: 57.2% (95% CI: 54.7– 59.6%), CLIP: 28.8% (95% CI: 26.2–31.3%) and Fine-tuned CLIP 26.8%, as detailed in Supplementary A.1.4). The analysis of unimodal input results and additional classification metrics further aligns with these findings. Specifically, embedding model inputs with expertise-curated knowledge significantly outperforms the use of general domain knowledge. These observations highlight the necessity for fine-tuning and domain-specific adaptation or embedding images and texts in medical applications. Furthermore, they affirm the efficacy of our MMCBM as a viable and effective means to achieve model interpretability without compromising algorithmic performance.

Evaluation of image-concept alignment

Our MMCBM demonstrates classification performance on par with state-of-the-art black-box models and offers interpretable insights into final model outputs. We have also shown that the quality of model interpretability depends on the quality of (1) the prior knowledge used to construct the MMCBM concept bank, (2) the image and concept embedding functions, and (3) image-concept alignment. We sought to evaluate our model's interpretability according to these three aspects.

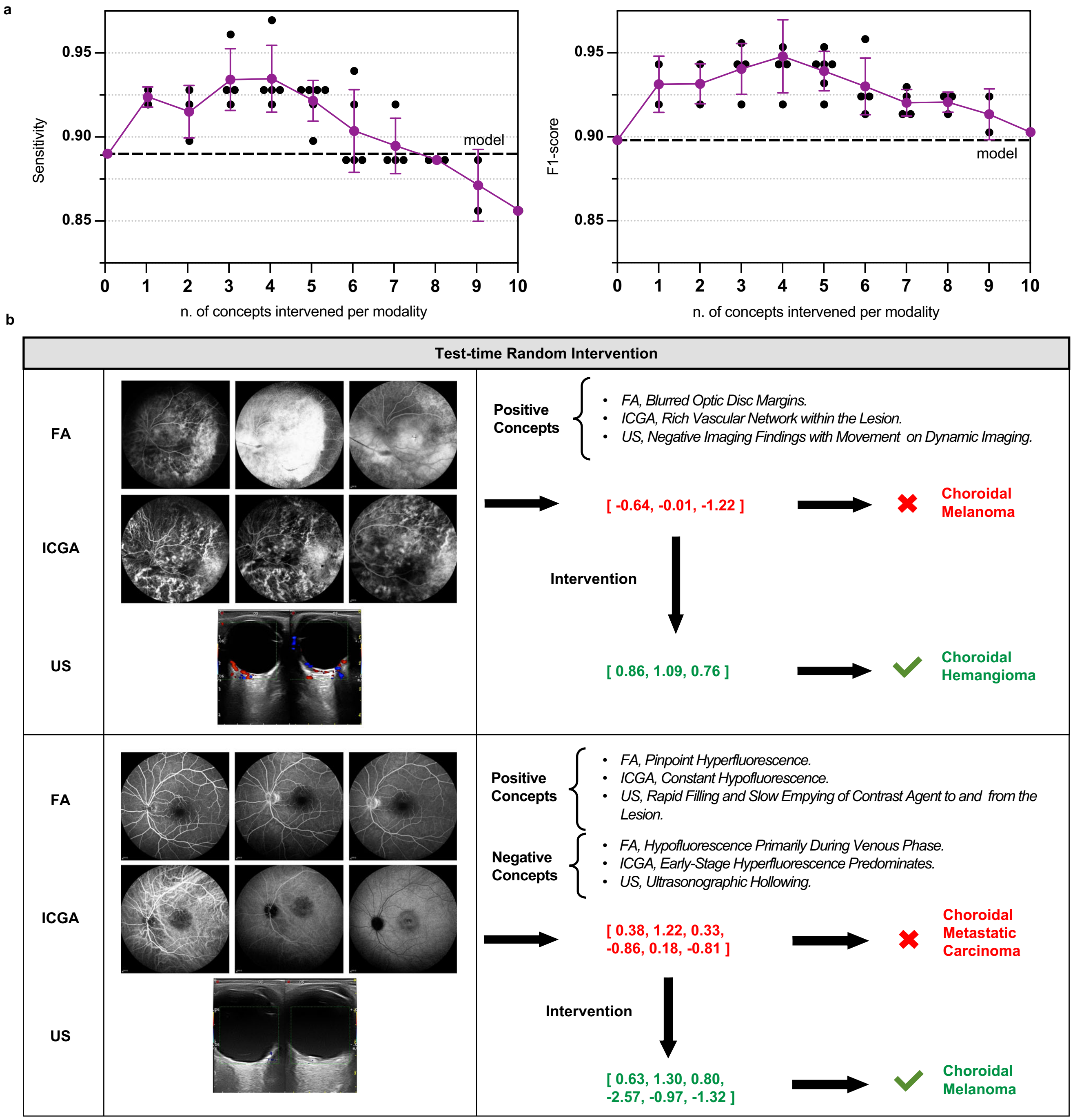


Fig. 5 | Comparison of the evaluation performance with and without AI assistance. a Performance benchmark with human evaluators: A comparison of our model's performance against inexperienced doctors (denoted as "Inexp") and senior doctors. After presenting them with the model's predicted concepts, they conducted a subsequent assessment, enabling us to document and compare performance metrics. Additionally, "MC/L" refers to model-generated concepts/labels, and "RC/L" refers to concepts or labels assigned randomly. Source data are provided as a Source Data file. **b** Qualitative examples of MMCBM assistance. Case 1: Initially misdiagnosed, the inexperienced doctors corrected the error after consulting the MMCBM's top ten predicted concepts, resulting in an accurate diagnosis. Case 2: Despite referencing the MMCBM's predicted concept, the inexperienced doctors maintained an erroneous diagnosis.

First, we evaluated the MMCBM feature representations and their accuracy in describing input images. Model representations for each of FA, ICGA, and US imaging studies were computed by the respective MMCBM encoders before leveraging t-SNE[40] dimensionality reduction techniques to visualize the complex feature landscapes from our multimodal dataset (Fig. 6a). We observe distinct clusters corresponding to hemangioma, metastatic carcinoma, and melanoma, indicating effective class separation by the MMCBM encoders. Qualitatively, the clusters corresponding to multimodal data inputs appear more cohesive and less dispersed, suggesting that integrating multimodal inputs may improve the separability of the different class representations in this representation space. This enhanced clustering density may contribute to the improved discriminative performance of our multimodal MMCBM models in contrast to models with only unimodal inputs accessible.

Next, we evaluated the quality of the MMCBM concept representation and image-concept alignment by examining the accuracy of the SVM classifiers employed in generating concept vectors for

Table 1 | Human performance with the AI assistant

	Human/model	Randomized part	Accuracy	Precision	Recall	F1-score
Blackbox	Model	None	93.1%	90.5%	87.8%	89.0%
	Human	None	89.7%	83.8%	88.3%	85.6%
		Label Only	62.7%	56.1%	63.0%	57.1%
MMCBM	Model	None	93.6%	91.5%	89.7%	90.5%
	Human	None	91.9%	86.2%	86.2%	86.2%
		Label Only	70.2%	61.1%	68.2%	63.2%
		Concept Only	90.3%	87.5%	85.9%	86.6%
		Concept & Label	59.1%	51.3%	56.9%	52.6%

For the "Randomized part", "None" means providing the real output from the model, Label only means that only the label is randomized, "Concept only" means that only the concept is randomized, and "Concept" and "Label" means that both the concept and label are randomized. Note: the attention map (GradCAM) is incorporated into the visualization of the black-box assistant.

Table 2 | Comparison of the evaluation performance with and without AI assistance

Init	Assist	Wrong then right	Right then wrong	Right then right	Wrong then wrong
N A	BB + ML	41.4%	3.5%	48.3%	6.9%
	MMCBM + MC + ML	43.1%	3.5%	48.3%	5.2%
	BB + RL	25.9%	13.8%	37.9%	22.4%
	MMCBM + MC + RL	32.8%	8.6%	43.1%	15.5%
BB +RL	MMCBM + MC + RL	22.4%	10.3%	53.5%	13.8%
	MMCBM + MC + ML	29.3%	1.7%	62.1%	6.9%

The group "Wrong then Right" means that the inexperienced evaluator made a mistake in the initial judgment yet corrected it with the provided assistance. The group "Right then Wrong", "Right then Right", and "Wrong then Wrong" are defined similarly. The stage "Init: NA" denotes that the evaluator initially performs the task with no assistance and the stage "Assist: BB + ML" denotes that the evaluator further performs the task with the assistance of black-box model with model-generated labels. For other notations, "MC" means model-generated concepts and "RL" means randomly generated labels.

each medical concept. A high SVM accuracy score indicates a concept's representational effectiveness and consistent presence across the dataset. According to this metric, FA and ICGA concepts achieve high accuracy across the board (Fig. 6b), with accuracy on test data exceeding 90% for all concepts. This suggests that concepts derived from FA and ICGA are well-represented and aligned with the input images. In contrast, though less accurate, the accuracy scores for US-based concepts are still higher than 80% for all concepts. This suggests that classifying diseases from ultrasound images alone may be more challenging. Specific details of the individual concepts and their corresponding accuracies are detailed in Supplementary Table 6.

To further assess the quality of MMCBM concept-based interpretability, we examined how well the model concepts align with ophthalmologist annotations. We selected the top-k concepts predicted by MMCBM for each patient in the multimodal testing dataset. In Fig. 6c, we quantify our model's alignment with expert annotations according to key performance metrics: Precision@k, Recall@k, F1@k, Median-Rank@k, Mean-Rank@k, and mean–reciprocal–rank-(MRR) @k, with $k = 10$. We compared two setups of concept banks: the report-extracted and the expert-verified. We found that report-extracted concepts achieved Precision@10 = 0.53 and Recall@10 = 0.57, similar to expert-verified concepts (Precision@10 = 0.54, Recall@10 = 0.55). It is worth noting that expert-verified concepts yielded better alignment with expert annotations, suggesting that human intervention in the verification process improves the concept bank's ability to capture domain knowledge. Our analysis demonstrates that the MMCBM model concepts extracted from reports closely match the performance of expert-verified annotations across various metrics. This suggests that report-extracted concepts achieve interpretability comparable to expert-verified concepts, negating the need for time-intensive expert annotation while effectively capturing the salient clinical features of interest to ophthalmologists.

Demonstration of human–model interaction

To exemplify a practical engineered system for enabling human–model interactions, we make available our website (https://mmcbm.liuy.site) used for this user-based study with the eight ophthalmologists. Our website provides a user-friendly online interface for concept bank verification and predication evaluation. The annotation system allows ophthalmologists to upload de-identified images and annotate them with clinically meaningful concepts (Fig. 7a) or verify images along with report-extracted concepts. The prediction system can accept FA, ICGA, and/or US images, and use them to output imaging concepts with confidence scores (Fig. 7b). In instances where MMCBM may produce erroneous concept predictions, clinicians can easily adjust the confidence scores of individual concepts within the user interface. Such adjustments can refine and correct model predictions to better align them with clinical findings that may be otherwise inaccessible to the model. This feature of human intervention significantly improves the practical utility of MMCBMs in clinical decision-making, fostering a more collaborative and accurate diagnostic process. Figure 7c displays several examples generated by MMCBM, including a curated selection of representative cases processed by the model. Finally, given the model outputs, a basic diagnostic report can be generated by leveraging LLMs to interpret the MMCBM outputs and concept activations (Fig. 7d and Supplementary Fig. 8). The generative model highlights the top-k activated concepts before presenting the final generated diagnostic report. The report generation prompt example is included in Supplementary Fig. 8.

In summary, our results highlight the MMCBM model as a promising tool for clinical decision support. While the model's predictions are accurate on their own, they are most effective when combined with human expertise, offering the most comprehensive diagnostic performance and underscoring the potential of AI-assisted diagnostics.

Discussion

In this work, we establish the Multimodal Medical Concept Bottleneck Model (MMCBM), an interpretable approach for diagnosing rare diseases. To facilitate the application of advanced machine learning techniques, we initially tackled the significant challenge of scarce comprehensive training data by curating the choroidal tri-modal imaging clinical dataset. This dataset, which includes image data of fluorescein angiography (FA), indocyanine green angiography (ICGA), and ultrasound (US) with associated radiology reports, to our knowledge, is an extensive dataset of choroidal melanoma. Based on this dataset, our MMCBM maintains the accuracy of prior "black-box" models and introduces interpretability through the concept bottleneck model. Furthermore, by incorporating the explainable MMCBM

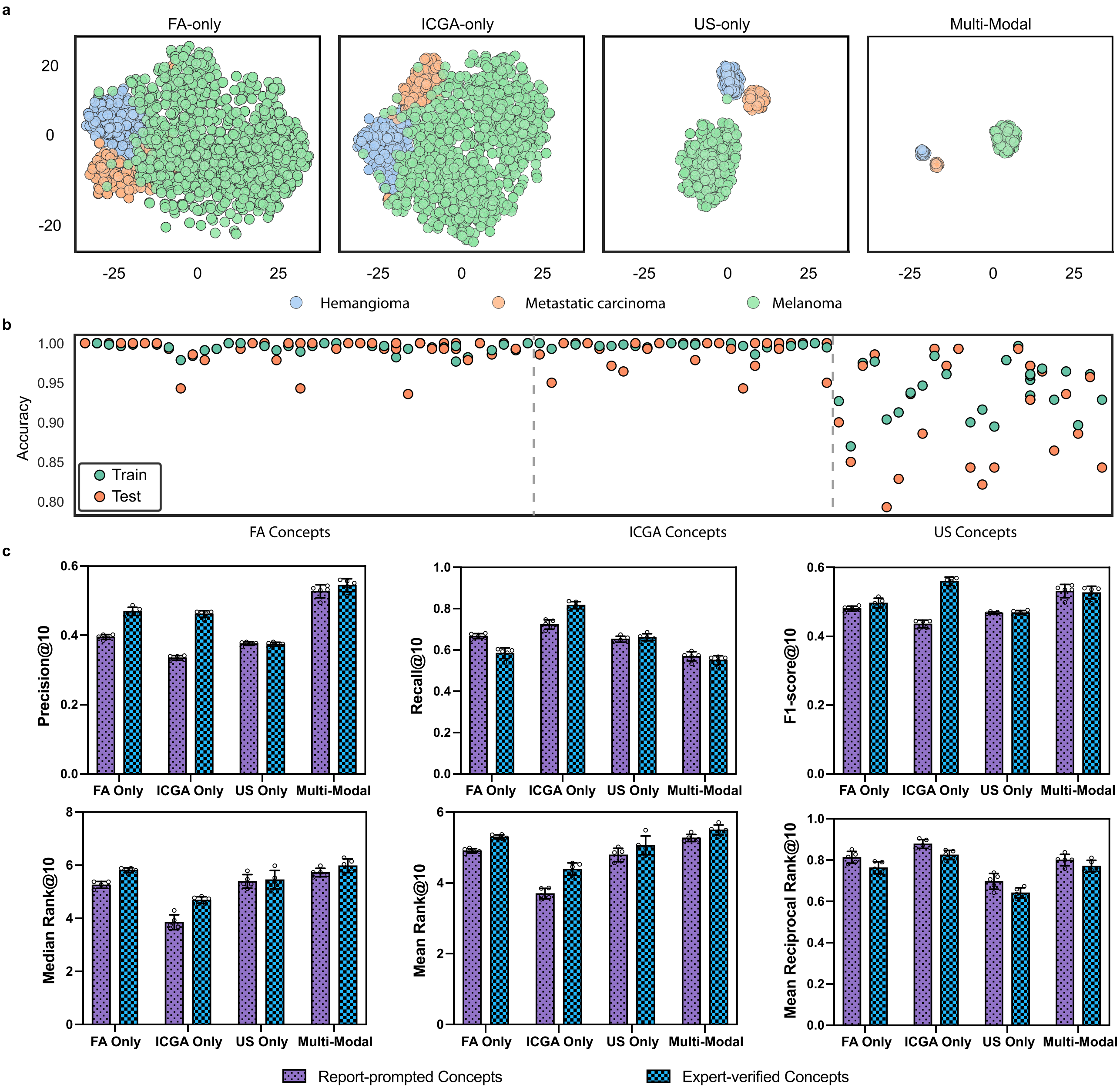
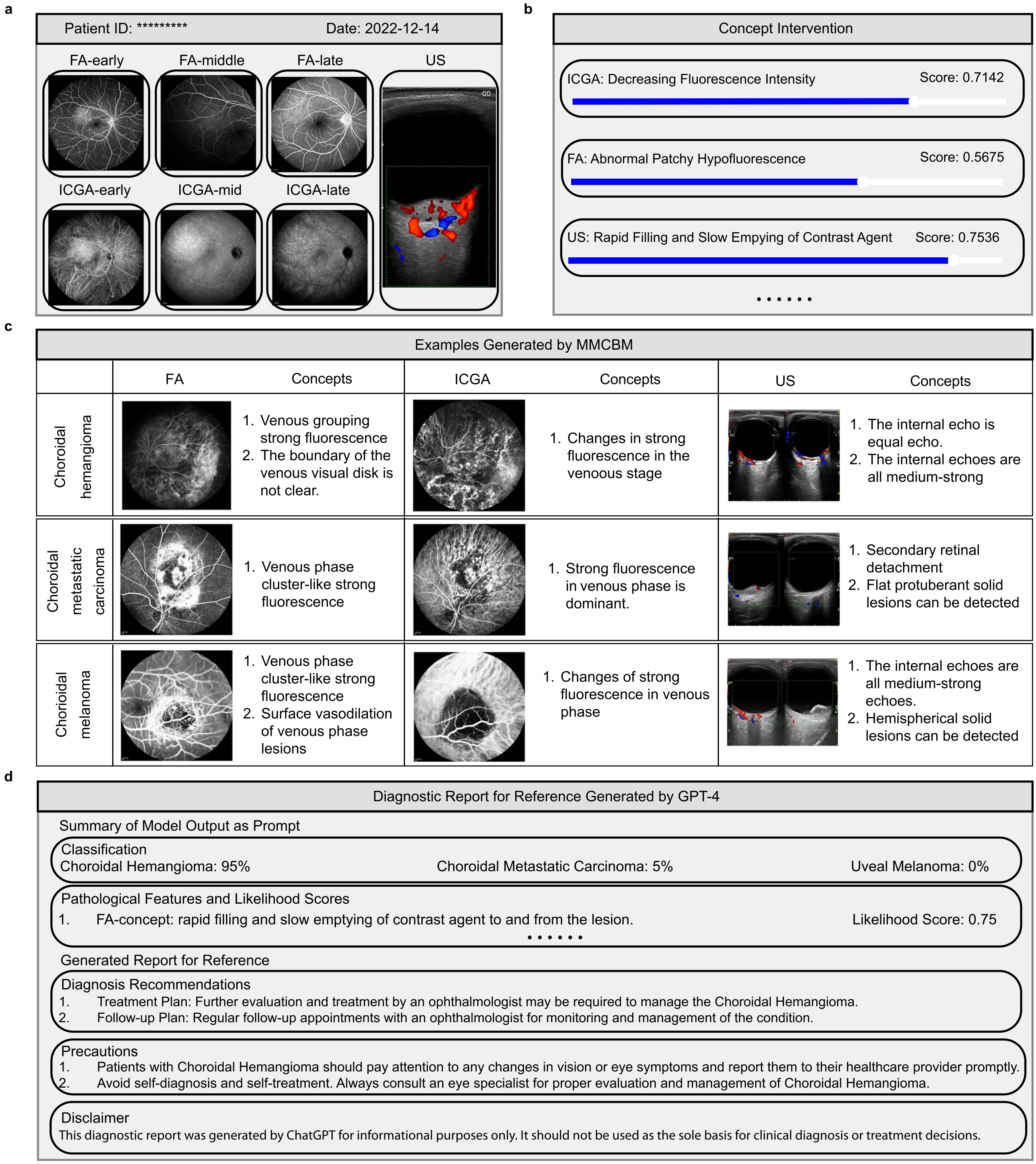


Fig. 6 | Comparative human evaluation and model insights. a Embedding Visualizations via t-SNE: This offers a graphical representation of embeddings from the trio of pretrained encoders. Notably, the fused MM embeddings are processed through the attention-pooling mechanism. **b** Accuracy of SVMs in generating concept banks using Concept Activation Vectors (CAVs). **c** Metrics of predicted Top-*k* concepts on test dataset with $k = 10$. This evaluation includes precision@*k*, recall@*k*, and F1@*k*, as well as mean rank@*k*, median rank@*k*, and mean reciprocal rank@*k*. The data were presented as mean ± SD, with error bars indicating the standard deviation from $n = 5$ independent replicates. Source data are provided as a Source Data file.

into the diagnostic workflow, our model significantly enhances the performance of inexperienced ophthalmologists.

Unlike traditional methods for explainable AI, which often rely on saliency maps[41–43] to highlight important spatial attributes, our pipeline aligns more closely with clinical practice by mimicking the diagnostic thought process used by domain experts. Clinicians identify a range of descriptive visual features, including textual elements, contrast, shape, and dynamic changes, that extend beyond pixel values alone. Our approach of constructing “concepts” extends its definition from the data-driven high-level visual features[44–46] to the image-context pairs so that the extracted image features align well with what the radiologist pays attention to in the clinical practice, thus providing human-comprehensible descriptions that facilitate intervention in the diagnostic process. This yields benefits in both methodology and clinical practice. Regarding the methodology, compared to the traditional approaches requiring either expensive labeling or sophisticated network infrastructure designs to integrate clinical insights, it simplifies the alignment between domain knowledge in clinical practice and the representational power of neural networks. In terms of clinical practice, it facilitates not only the radiologists in identifying overlooked biomarkers by reminding them of the image-concept pairs but also the ophthalmologists in the diagnosis by demonstrating the decision routine with concept evidence. Our results prove immensely beneficial for inexperienced doctors who may lack training in finding identification and risk over-reliance on AI outputs[47].

Moreover, recent advancements in vision and natural language processing, such as large language models (LLMs) and contrastive language-image pre-training (CLIP), have paved new pathways for



Concept Intervention

ICGA: Decreasing Fluorescence Intensity — Score: 0.7142

FA: Abnormal Patchy Hypofluorescence — Score: 0.5675

US: Rapid Filling and Slow Empying of Contrast Agent — Score: 0.7536

……

Examples Generated by MMCBM

	FA	Concepts	ICGA	Concepts	US	Concepts
Choroidal hemangioma		1. Venous grouping strong fluorescence 2. The boundary of the venous visual disk is not clear.		1. Changes in strong fluorescence in the venoous stage		1. The internal echo is equal echo. 2. The internal echoes are all medium-strong
Choroidal metastatic carcinoma		1. Venous phase cluster-like strong fluorescence		1. Strong fluorescence in venous phase is dominant.		1. Secondary retinal detachment 2. Flat protuberant solid lesions can be detected
Choroidal melanoma		1. Venous phase cluster-like strong fluorescence 2. Surface vasodilation of venous phase lesions		1. Changes of strong fluorescence in venous phase		1. The internal echoes are all medium-strong echoes. 2. Hemispherical solid lesions can be detected

Diagnostic Report for Reference Generated by GPT-4

Summary of Model Output as Prompt

Classification
Choroidal Hemangioma: 95% Choroidal Metastatic Carcinoma: 5% Uveal Melanoma: 0%

Pathological Features and Likelihood Scores
1. FA-concept: rapid filling and slow emptying of contrast agent to and from the lesion. Likelihood Score: 0.75
……

Generated Report for Reference

Diagnosis Recommendations
1. Treatment Plan: Further evaluation and treatment by an ophthalmologist may be required to manage the Choroidal Hemangioma.
2. Follow-up Plan: Regular follow-up appointments with an ophthalmologist for monitoring and management of the condition.

Precautions
1. Patients with Choroidal Hemangioma should pay attention to any changes in vision or eye symptoms and report them to their healthcare provider promptly.
2. Avoid self-diagnosis and self-treatment. Always consult an eye specialist for proper evaluation and management of Choroidal Hemangioma.

Disclaimer
This diagnostic report was generated by ChatGPT for informational purposes only. It should not be used as the sole basis for clinical diagnosis or treatment decisions.

Fig. 7 | **Example of human interactive interface.** We offer a website to facilitate the user interactive study with ophthalmologists and our trained MMCBM model. **a** Image display panel: as FA and ICGA imaging span various time frames, ophthalmologists pinpoint images from early, middle, and late phases for accurate classification. **b** Interventions interface on concept bottleneck: a panel that allows adjustment of the concept scores to refine the final prediction. **c** Visual emphasis on bottlenecks: a curated selection of representative cases processed by the model, highlighting the top-k concepts prioritized by their attention scores in the weight matrix displayed across three distinct tumor classes. **d** Diagnostic reporting in action: an example of a diagnostic report formulated by ChatGPT during the testing phase. The input to ChatGPT includes the predicted top-k concepts combined with patient-specific details, highlighting the model's capability to produce interpretable diagnoses.

research into interpretable diagnostic systems. However, for rare diseases like choroidal melanoma, the scarcity of paired image-text knowledge on the internet presents a significant challenge to the reliability of these models' reasoning capabilities, as evidenced in Fig. 3. While professional annotation of high-quality data can mitigate this issue, further data access and expertise challenges remain[38], especially for rare diseases. Our concept-based multimodal model circumvents these challenges by utilizing LLMs to process texts without necessitating detailed labeling of image features. The model's predictive and interpretive power stems from integrating the pretrained model with the extracted relationship between reports and images. This approach mitigates the data scarcity issue for rare diseases in recent foundation models, avoiding the need for extensive labeling efforts in medical AI preparation, thus making the design extendable to other rare diseases.

In the realism of AI-aid medical diagnosis, particularly for the detection and intervention of serious diseases like the choroid neoplasias we considered in the current work, ethical considerations are of critical importance[48]. Our methodology, which enables human-in-the-loop feedback, helps address this issue by aligning human expertise with AI diagnosis. Specifically, by actively involving domain experts in the training and validation phases of AI model development, we not only ensure that the AI's diagnostic concepts are vetted by experienced clinicians but also provide feasible constraints of the degree of AI intervention. This reduces the risk of hallucinations arising from reliance solely on AI. This approach may foster trust among clinicians and patients in AI-assisted medical decisions. The inclusion of human-in-the-loop integration in our AI models aligns with ethical guidelines for AI in healthcare, emphasizing the safeguarding of patient dignity and privacy. As we advance the frontiers of medical AI, it is crucial to maintain a balanced synergy between technological innovation and ethical responsibility, ensuring that AI serves as a supportive tool rather than a replacement for the nuanced judgment of medical professionals.

While the proposed MMCBM demonstrated improved generalization compared to black-box models, achieving broader generalizability will likely require multi-center collaborative efforts. Additionally, incorporating a wider range of tumor imaging modalities and medical reports could enhance the applicability of concept-based models to the diagnosis of other tumors with similar recognizable features. Furthermore, exploring how AI models can be integrated into clinical workflows will be crucial for advancing medical AI, particularly as more sophisticated AI tools continue to emerge.

In summary, the development of MMCBMs marks a significant advancement toward achieving interpretable and reliable diagnoses within the healthcare domain. As efforts to refine and incorporate these models into clinical workflows progress, it is imperative to carefully consider the ethical and regulatory dimensions to ensure that these innovations enhance patient outcomes without compromising the standards of care or jeopardizing patient safety. This work delineates a promising avenue for applying artificial intelligence in the nuanced and critical field of diagnosing rare diseases, offering a blueprint for future explorations in this vital area of medical research.

Methods

Dataset collection and ethics statement

All studies were conducted in accordance with protocols approved by the Ethics Committee of Beijing Tongren Hospital, Capital Medical University (Protocol No. TRECKY2018-056-GZ(2022)-07). The patient data in the CTI dataset were collected at Beijing Tongren Hospital from March 2013 to September 2019. Sex and/or gender were not considered in the study design; participants' sex/gender, race, ethnicity, and ancestry were determined through self-report. To our knowledge, it is a substantial clinical database containing multimodal data from patients with choroidal melanoma and other closely related ocular pathologies. This extensive database contains diagnostic and pathological data of patients with choroidal diseases. The database includes a total of 925 cases, which comprise 161 cases of choroidal hemangioma, 82 cases of choroidal metastatic carcinoma, and 682 cases of choroidal melanoma. The image collection includes three types of radiological images: fluorescein angiography (FA), indocyanine green angiography (ICGA), and Doppler ultrasound images (US). Each patient has one or more modalities of images. The FA and ICGA images, being time-series, were captured from three angles: 30, 55, and 102 degrees. The US images include two types: B-mode ultrasound and color Doppler ultrasound. Medical professionals have thoroughly reviewed the data-cleaning process to ensure its integrity and clinical relevance. For FA and ICGA modalities, we ignored the shooting angle and categorized the FA and ICGA images into three periods—early, middle, and late—in alignment with existing clinical diagnostic recommendations. The time frames for these periods are as follows: ICGA (Early: less than 5 min; Middle: between 5 and 20 min; and Late: at least 20 min) and FA (Early: less than 5 min, Middle: between 5 and 10 min, Late: at least 10 min). We selected binocular color Doppler images containing blood flow information for the US modality. Finally, the cleaned dataset includes a total of 750 cases, which comprises 128 cases of choroidal hemangioma, 80 cases of choroidal metastatic carcinoma, and 542 cases of choroidal melanoma. There are 53 patients with choroidal hemangioma, 38 patients with choroidal metastatic carcinoma, and 194 patients with choroidal melanoma with all three imaging modalities, which we refer to as multimodal (MM) data. Additionally, 97 cases have clinical diagnostic reports that describe the radiological features observed in the FA, ICGA, and US images. Informed consent was obtained from all patients whose anonymized and de-identified data is included in the dataset. Per the Declaration of Helsinki 2000, the collecting organization obtained written informed consent from the patients.

Data splitting

To optimize data utilization and establish reliable evaluation indicators, we initially allocated 20% of patients with all three imaging studies as the test set and performed 5-fold cross-validation at the patient level on the remaining data. Specifically, the remaining data is split into five folds based on each pathology and modality. Data augmentation techniques were applied during training, including random horizontal flipping, random rotating, and random zooming. To build the multimodal concept banks, we used 97 diagnosis reports, comprising 39 cases of choroidal hemangioma, 18 of choroidal metastatic carcinoma, and 40 of choroidal melanoma. Each report included three modal images and prompted GPT-4 to extract relevant medical concepts from reports. The prompts are detailed in Supplementary Fig. 7, and the extracted concepts are in Supplementary Table 6.

Model training

Consider a training dataset $\mathcal{D}_{train} = \{(\mathbf{x}, \mathbf{r}, y)\}$ comprising image-report pairs, where $\mathbf{x} \in \mathcal{X}$ represents a fundus image (of any imaging modality), $\mathbf{r} \in \mathcal{R}$ is the clinical patient report collected by doctors, $y \in \mathcal{Y} := \{\text{hemangioma, carcinoma, melanoma}\}$ is the corresponding disease label. We utilize GPT-4 to analyze the reports and extract relevant concepts, represented as a function $\text{LLM} : \mathcal{R} \rightarrow \mathcal{C}$ where $\mathcal{C}$ is the space of concepts. We can then prompt GPT-4 to combine concepts with the same semantic meaning, resulting in a compressed representation of N concepts $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$. Using a pretrained multi-modality backbone $\phi : \mathcal{X} \rightarrow \mathcal{Z}$ capable of mapping different modality images into a shared feature space, we can generate bottleneck embeddings to establish a concept bank, denoted as $\mathcal{Z}_{\mathcal{C}} \in \mathbb{R}^{N \times d}$, where N is the number of concepts and d the size of the embedding space of ϕ. Row i of the two-dimensional matrix $\mathcal{Z}_{\mathcal{C}}$ represents the learned representation of the ith concept c_i obtained through Concept Activation Vectors (CAVs)[35]. MMCBM generates a prediction $\hat{y} = g(\text{sim}(\phi(\mathrm{x}), \mathcal{Z}_{\mathcal{C}}))$. The function $\text{sim} : \mathbb{R}^{\mathrm{d}} \rightarrow \mathbb{R}^{\mathrm{N}}$ computes the concept scores by calculating the similarities between image features and each element of the concept bank $\mathcal{Z}_{\mathcal{C}}$. The function $g : \mathbb{R}^N \rightarrow \mathcal{Y}$ predicts the final label based on the concept scores, serving as an interpretable predictor. To learn the MMCBM, we solve the following problem:

$$\min_{g} \underset{(\mathbf{x}, c, y) \sim \mathcal{D}}{\mathbb{E}} \mathcal{L}[g(\text{sim}(\phi(x), \mathcal{Z}_{\mathcal{C}})), y] \tag{1}$$

where $\phi(x)$ is the projection to the concept space and $\mathcal{L}$ is the cross-entropy loss. To ensure that the final prediction $\hat{y}$ can be easily derived from input $\text{sim}(\phi(x), \mathcal{Z}_{\mathcal{C}})$, we model g as a linear classifier.

Enhancing diagnostic with LLM

GPT-4 was employed in two stages of the workflow. For model training, we utilize GPT-4 to analyze clinical reports and extract relevant concepts for use in training the MMCBM model. Additionally, GPT-4 was used for medical report generation (MRG). By combining the predicted concepts with the model's output, we prompted GPT-4 to generate comprehensive clinical reports (Supplementary Fig. 8). These reports follow a structured format, including patient information, medical details, diagnosis, and treatment recommendations. GPT-4 played a key role in converting the extracted concepts into a cohesive, readable narrative, ensuring the generation of standardized clinical reports.

Evaluation of model performance

Using a fivefold cross-validation framework, we report the macro-averaged metrics accuracy, precision, recall, and F1 score, which considers both precision and recall while addressing potential class imbalances. In addition to these traditional classification metrics, we also focused on interpretability metrics such as Precision@k, Recall@k, Mean Rank@k, and Median Rank@k. Precision@k measures how many of the top-k identified concepts were right compared with the annotated ground truth. Recall@k evaluates the ratio of correct concepts in the first k predictions to all correct concepts for the patient. F_1@k is the harmonic mean of Precision@k and Recall@k. Mean Rank@k and Median Rank@k indicate the average ranking position of the correct concept; lower scores are better.

Statistical information

For the comparison of model performance in Fig. 3b, the error bar is defined as the standard error and significance is calculated through the two-sample t-tests based on the distribution of metrics obtained from the k-folds. The *n.d.* denotes no difference, which indicates the p-value associated with the test is larger than 0.05. For the comparison of model performance in Figs. 3d, 5a and Table 1, we bootstrap the test dataset with the leave-one-out setup and calculate the statistical significance with the two-sample proportions z-tests.

Software utilized

All code was implemented in Python (3.11) using Pytorch (2.0.1) as the base deep learning framework. We also used several Python packages for data analysis and results visualization, including monai (1.3.2), openai (1.44.0), torchvision (0.15.2), numpy (1.24.4), scikit-learn (1.3.0), pandas (2.0.3), matplotlib (3.8.2), opencv-python (4.8.0), and gradio (4.43.0). Prism was used to create Figs. 2, 3, 5, 4, 6.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The Choroid Tri-modal Imaging (CTI)[49] dataset utilized in this study is available in figshare with the identifier https://doi.org/10.6084/m9.figshare.28255265.v2. The raw patient data were not publicly available due to patient privacy restrictions. Additionally, source data for figures are provided with this paper in the Source Data file. Source data are provided with this paper.

Code availability

The code is publicly available under the BSD License at https://github.com/brain-intelligence-lab/MMCBM. A permanent version is released on Zenodo[50].

References

1. Chan, H.-P., Hadjiiski, L. M. & Samala, R. K. Computer-aided diagnosis in the era of deep learning. *Med. Phys.* **47**, e218–e227 (2020).
2. Johnson, A. E. et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**, 317 (2019).
3. Lin, M. et al. Improving model fairness in image-based computer-aided diagnosis. *Nat. Commun.* **14**, 6261 (2023).
4. Gao, M. et al. Discriminative ensemble meta-learning with co-regularization for rare fundus diseases diagnosis. *Med. Image Anal.* **89**, 102884 (2023).
5. Menze, B. H. et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Trans. Med. Imaging* **34**, 1993–2024 (2014).
6. Liew, S.-L. et al. A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. *Sci. Data* **5**, 1–11 (2018).
7. Gatta, G. et al. Rare cancers are not so rare: the rare cancer burden in europe. *Eur. J. Cancer* **47**, 2493–2511 (2011).
8. Decherchi, S., Pedrini, E., Mordenti, M., Cavalli, A. & Sangiorgi, L. Opportunities and challenges for machine learning in rare diseases. *Front. Med.* **8**, 747612 (2021).
9. Molnar, M. J. & Molnar, V. Ai-based tools for the diagnosis and treatment of rare neurological disorders. *Nat. Rev. Neurol.* **19**, 455–456 (2023).
10. Wang, S.-H. et al. Global development of artificial intelligence in cancer field: a bibliometric analysis range from 1983 to 2022. *Front. Oncol.* **13**, 1215729 (2023).
11. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T. J. & Zou, J. A visual–language foundation model for pathology image analysis using medical twitter. *Nat. Med.* **29**, 2307–2316 (2023).
12. Zhang, X., Wu, C., Zhang, Y., Xie, W. & Wang, Y. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nat. Commun.* **14**, 4542 (2023).
13. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. 38th International Conference on Machine Learning*, (*PMLR*) 8748–8763 (2021).
14. Chae, A. et al. Strategies for implementing machine learning algorithms in the clinical practice of radiology. *Radiology* **310**, e223170 (2024).
15. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. Ai in health and medicine. *Nat. Med.* **28**, 31–38 (2022).
16. Varoquaux, G. & Cheplygina, V. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digit. Med.* **5**, 48 (2022).
17. Yu, F. et al. Heterogeneity and predictors of the effects of ai assistance on radiologists. *Nat. Med.* **30**, 837–849 (2024).
18. Koh, P. W. et al. Concept bottleneck models. In *Proc. 37th International Conference on Machine Learning (PMLR)* 5338–5348 (2020).
19. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl Acad. Sci. USA* **116**, 22071–22080 (2019).
20. Jager, M. J. et al. Uveal melanoma. *Nat. Rev. Dis. Prim.* **6**, 24 (2020).
21. Shields, C. L. et al. Metastatic tumours to the eye. review of metastasis to the iris, ciliary body, choroid, retina, optic disc, vitreous, and/or lens capsule. *Eye* **37**, 809–814 (2023).
22. Kaliki, S., Shields, C. L. & Shields, J. A. Uveal melanoma: estimating prognosis. *Indian J. Ophthalmol.* **63**, 93 (2015).
23. Luo, J. et al. Characteristics, treatments, and survival of uveal melanoma: a comparison between chinese and american cohorts. *Cancers* **14**, 3960 (2022).
24. Singh, A. D., Turell, M. E. & Topham, A. K. Uveal melanoma: trends in incidence, treatment, and survival. *Ophthalmology* **118**, 1881–1885 (2011).
25. Mathis, T. et al. New concepts in the diagnosis and management of choroidal metastases. *Prog. Retin. Eye Res.* **68**, 144–176 (2019).
26. Ajani, J. A. et al. Gastric cancer, version 2.2022, nccn clinical practice guidelines in oncology. *J. Natl Compr. Cancer Netw.* **20**, 167–192 (2022).

27. Egan, K. M., Seddon, J. M., Glynn, R. J., Gragoudas, E. S. & Albert, D. M. Epidemiologic aspects of uveal melanoma. *Surv. Ophthalmol.* **32**, 239–251 (1988).
28. Augsburger, J. J. & Gamel, J. W. Clinical prognostic factors in patients with posterior uveal malignant melanoma. *Cancer* **66**, 1596–1600 (1990).
29. Carvajal, R. D. et al. Metastatic disease from uveal melanoma: treatment options and future prospects. *Br. J. Ophthalmol.* **101**, 38–44 (2016).
30. Khoja, L. et al. Meta-analysis in metastatic uveal melanoma to determine progression free and overall survival benchmarks: an international rare cancers initiative (irci) ocular melanoma study. *Ann. Oncol.* **30**, 1370–1380 (2019).
31. Tan, M. & Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proc. 36th International Conference on Machine Learning* (*PMLR*) 6105–6114 (2019).
32. Chen, R. J. et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell* **40**, 865–878 (2022).
33. Safari, P., India, M. & Hernando, J. Self-attention encoding and pooling for speaker recognition. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association: Virtual Event, Shanghai, China* 941–945 (International Speech Communication Association (ISCA), 2020).
34. Achiam, J. et al. Gpt-4 technical report. Preprint at arXiv:2303.08774 (2023).
35. Kim, B. et al. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (tcav). In *Proc. 35th International conference on machine learning (PMLR)* 2668–2677 (2018).
36. Yuksekgonul, M., Wang, M. & Zou, J. Post-hoc concept bottleneck models. *International conference on learning representations (ICLR)* (2023).
37. Yang, Y. et al. Language in a bottle: language model guided concept bottlenecks for interpretable image classification. *In Proc. IEEE Conference on Computer Vision and Pattern Recognition* 19187–19197 (2023).
38. Wang, Z., Wu, Z., Agarwal, D. & Sun, J. Medclip: Contrastive learning from unpaired medical images and text. In *Proc. of the Conference on Empirical Methods in Natural Language Processing* Vol. 2022 3876 (2022).
39. Zhang, S. et al. Large-scale domain-specific pretraining for biomedical vision-language processing. Preprint at arXiv:2303.00915 (2023).
40. Van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *J. Machine Learn.Res.* **9**, 2579–2605 (2008).
41. Selvaraju, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. *In Proc. of the IEEE international conference on computer vision* 618–626 (2017).
42. Wu, Y. et al. Interpretable identification of interstitial lung disease (ILD) associated findings from CT. In *Medical Image Computing and Computer Assisted Intervention* 560–569 (Springer, 2020).
43. Tjoa, E. & Guan, C. A survey on explainable artificial intelligence (xai): toward medical xai. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 4793–4813 (2020).
44. Sauter, D. et al. Validating automatic concept-based explanations for ai-based digital histopathology. *Sensors* **22**, 5346 (2022).
45. Lucieri, A. et al. Exaid: A multimodal explanation framework for computer-aided diagnosis of skin lesions. *Comput. Methods Prog. Biomed.* **215**, 106620 (2022).
46. Pinckaers, H. et al. Predicting biochemical recurrence of prostate cancer with artificial intelligence. *Commun. Med.* **2**, 64 (2022).
47. Kostick-Quenet, K. M. & Gerke, S. Ai in the hands of imperfect users. *npj Digit. Med.* **5**, 197 (2022).
48. Char, D. S., Abràmoff, M. D. & Feudtner, C. Identifying ethical considerations for machine learning healthcare applications. *Am. J. Bioeth.* **20**, 7–17 (2020).
49. Yang, X., Liu, Y., Wu, Y., Wei, W. & Gu, S. CTI Dataset. figshare. https://doi.org/10.6084/m9.figshare.28255265.v2 (2025).
50. Liu, Y. & Wu, Y. A Concept-based interpretable model for the diagnosis of choroid neoplasias using multimodal data. *Zenodo* https://doi.org/10.5281/zenodo.14774298 (2025).

Acknowledgements

This work was supported by NSFC Key Program (Grant No. 62236009 to S.G.), NSFC Program (Grant No. 82220108017, 82141128, and 82002883 to X.Y. and W.W.), the Capital Health Research and Development of Special (Grant No. 2024-1-2052 to X.Y. and W.W.), Science & Technology Project of Beijing Municipal Science & Technology Commission (Grant No. Z201100005520045 to X.Y. and W.W.), Sichuan Natural Science Fund for Creative Research Groups (Grant No. 2024NSFTD0041), and Sanming Project of Medicine in Shenzhen (Grant No. SZSM202311018 to X.Y. and W.W.).

Author contributions

Y.W., W.W., and S.G. designed and conceptualized the project. Y.W., Y.L., X.Y., and S.G. carried out the main analyses. Y.W., Y.L., Y.Y., M.S.Y., and S.G. contributed to the methodology. W.Y., X.S., L.Y., D.L., YM.L., S.Y., C.L., M.Z., X.Y., and W.W. curated the data collection and performed the data quality assessment. Y.W., Y.L., M.S.Y., J.C.G., and S.G wrote the manuscript. All authors discussed and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-58801-7.

Correspondence and requests for materials should be addressed to Xuan Yang, Wenbin Wei or Shi Gu.

Peer review information *Nature Communications* thanks Kaustav Bera, who co-reviewed with Palak Gupta, Martine Jager and Haotian Lin for their contribution to the peer review of this work. A peer review file is available.

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

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