npj Digital Medicine

https://doi.org/10.1038/s41746-025-02300-y

Article in Press

Grounded report generation for enhancing ophthalmic ultrasound interpretation using Vision-Language Segmentation models

Received: 11 June 2025

Accepted: 17 December 2025

Kai Jin, Qixuan Sun, Daohuan Kang, Ziyao Luo, Tao Yu, Wenzheng Han, Yi Zhang, Meng Wang, Danli Shi & Andrzej Grzybowski

Cite this article as: Jin, K., Sun, Q., Kang, D. *et al.* Grounded report generation for enhancing ophthalmic ultrasound interpretation using Vision-Language Segmentation models. *npj Digit. Med.* (2025). https://doi.org/10.1038/s41746-025-02300-y

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Grounded Report Generation for Enhancing Ophthalmic Ultrasound Interpretation using Vision-Language Segmentation Models

Kai Jin [a,b]*, Qixuan Sun [c], Daohuan Kang [d], Ziyao Luo [a,b], Tao Yu [a,b], Wenzheng Han [e], Yi Zhang [f], Meng Wang [g,h], Danli Shi [i,j], Andrzej Grzybowski [k,l].

[a] *Eye Center of Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China.*

[b] *Zhejiang Provincial Key Laboratory of Ophthalmology, Zhejiang Provincial Clinical Research Center for Eye Diseases, Zhejiang Provincial Engineering Institute on Eye Diseases, Hangzhou, China.*

[c] *Department of Biomedical Engineering, Zhejiang University, Hangzhou, China.*

[d] *Department of Ophthalmology, Children's Hospital, Zhejiang University School of Medicine，National Clinical Research Center for Child Health, Hangzhou, China.*

[e] *The First Affiliated Hospital, Wannan Medical College, Wuhu, Anhui, China.*

[f] *Department of Ophthalmology, the First Affiliated Hospital of Zhejiang Chinese Medicine University, Hangzhou, China.*

[g] *Centre for Innovation and Precision Eye Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore.*

[h] *Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore.*

[i] *School of Optometry, The Hong Kong Polytechnic University, Kowloon, Hong Kong.*

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

[k] *Institute for Research in Ophthalmology, Foundation for Ophthalmology Development, Poznan, Poland.*

[l] *Department of Ophthalmology, University of Warmia and Mazury, Olsztyn, Poland.*

*Corresponding author: Kai Jin, jinkai@zju.edu.cn.

Abstract

Accurate interpretation of ophthalmic ultrasound is crucial for diagnosing eye conditions but remains time-consuming and requires significant expertise. With the increasing volume of ultrasound data, there is a need for Artificial Intelligence (AI) systems capable of efficiently analyzing images and generating reports. Traditional AI models for report generation cannot simultaneously identify lesions and lack interpretability. This study proposes the Vision-Language Segmentation (VLS) model, combining Vision-Language Model (VLM) with the Segment Anything Model (SAM) to improve interpretability in ophthalmic ultrasound imaging. Using data from three hospitals, totaling 64,098 images and 21,355 reports, the VLS model achieved a BLEU4 score of 66.37 in internal test set, and 85.36 and 73.77 in external test sets. The model achieved a mean dice coefficient of 59.6% in internal test set, and dice coefficients of 50.2% and 51.5% with specificity values of 97.8% and 97.7% in external test sets, respectively. Overall diagnostic accuracy was 90.59% in internal and 71.87% in external test sets. A cost-effectiveness analysis demonstrated a 30-fold reduction in report costs, from $39 per report by senior ophthalmologists to $1.3 for VLS. This approach enhances diagnostic accuracy, reduces manual effort, and accelerates workflows, offering a promising solution for ophthalmic ultrasound interpretation.

Introduction

Accurate interpretation of medical images and the generation of comprehensive narrative reports are crucial for patient care, yet they place considerable demands on clinical professionals[1]. In ophthalmology, ophthalmic ultrasound plays a pivotal role, offering invaluable insights for diagnosing and managing a wide array of eye conditions, such as retinal diseases, and ocular tumors[2,3]. As a non-invasive imaging technique, it provides clinicians with detailed structural information that guides treatment decisions and disease monitoring. However, interpreting ophthalmic ultrasound images remains a time-intensive task and requires substantial expertise[4,5]. The growing volume of ultrasound data in clinical settings further complicates this process, increasing the need for advanced systems capable of analyzing the images and generating meaningful diagnostic reports[6].

While traditional AI models have shown substantial progress in medical image analysis and screening, their integration with report generation remains limited[7,8]. Many existing models lack interpretability, producing automated outputs that often require further examination and explanation[9,10]. Moreover, they often fail to highlight key lesions or abnormalities, which significantly reduces their practical utility. Furthermore, most AI tools are confined to specific medical specialties, limiting their broader applicability[11]. Previous research has shown that while the deep learning models achieves high accuracy in automating the classification of ophthalmic diseases using ultrasound images, it is limited to disease screening and lacks the ability to generate detailed image reports or provide precise clinical interpretations[12-14]. As a result, there is a notable gap in AI systems that combine image analysis with clear, interpretable reports, making them useful across various domains in medicine.

Recent advancements in artificial intelligence (AI), particularly through Vision-Language Model (VLM), are transforming ophthalmic diagnostics[15]. These models integrate visual and textual data to enhance diagnostic accuracy and support clinical decision-making by interpreting complex ocular images and generating comprehensive reports[16,17]. Meanwhile, Segment Anything Model (SAM), initially developed for natural images, has been adapted for medical imaging, enabling zero-shot segmentation of anatomical structures[18,19]. Combining VLM and SAM holds significant promise for improving diagnostic precision, and supporting personalized care. However, challenges such as ensuring model interpretability and reliability in clinical settings remain, requiring further research and development to fully integrate these technologies into ophthalmic care.

This study introduces a novel AI model that leverages advanced VLM to generate comprehensive diagnostic reports and annotate lesions directly on medical images. By merging image understanding with natural language processing, our approach creates grounded, meaningful reports, providing a scalable solution that integrates image analysis with explanatory report generation. Central to our approach is the Visual-Language Segmentation (VLS) model, which combines visual understanding with natural language processing, and the use of the SAM for precise lesion segmentation. These technologies enable the model to not only classify conditions but also generate accurate annotations on images. The model's potential extends beyond ophthalmology, offering a pathway for advancements in AI-driven diagnostic tools applicable to various medical imaging modalities, thus revolutionizing diagnostics across multiple medical specialties.

Results

This study utilized data from three distinct datasets: the Second Affiliated Hospital of Zhejiang University (SAHZU), the First Affiliated Hospital of Zhejiang Chinese Medical University (FAHZC), and the First Affiliated Hospital of Wannan Medical College (FAHWM), encompassing a total of 63,979 images and 21,239 real-world reports across seven ocular conditions. The demographic characteristics and distribution of findings from the datasets are detailed in **Table 1**. The overall workflow of the study is illustrated in **Figure 1**. Firstly, we conducted an automatic evaluation by comparing the performance of the developed VLS model with that of the VL model, as well as pre- and post-fine-tuned performance, to validate the strong capabilities of the VLS model in generating ocular ultrasound reports. Secondly, a systematic clinical effectiveness evaluation was performed by inviting one senior ophthalmologist and one junior ophthalmologist to assess report generation and diagnosis. Thirdly, we found that AI-assisted ocular ultrasound reporting demonstrated higher diagnostic accuracy and significantly reduced reporting time, validating the potential of our model as an auxiliary tool.

Demographic Data

Table 1 provides a comprehensive summary of the datasets used in this study, including details of the SAHZU datasets, as well as the FAHWM and FAHZC datasets used as external test sets. The data encompasses various patient, image, report, and diagnosis characteristics, and the following descriptions highlight key aspects: The training dataset consists of 5,497 patients, with 37,917 images and 12,649 reports, while the validation dataset includes 1,915 patients, 12,639 images, and 4,197 reports. The test dataset has 1,919 patients, 12,640 images, and 4,170 reports. Additionally, two external test sets are included: External Test Set 1 (FAHWM Dataset) contains 269 patients, 742 images, and 269 reports, while External Test Set 2 (FAHZC Dataset) consists of 70 patients, 160 images, and 70 reports. In total, 9,670 patients, 64,098 images, and 21,355 reports are included across all datasets. The mean age of patients across the datasets is similar, with a range of 49.5 to 49.7 years for the internal datasets and 50.8 years for External Test Set 1, and 57.4 years for External Test Set 2. The gender distribution is fairly balanced across the internal datasets, with 47.4% male and 52.6% female, while External Test Set 1 has a higher proportion of females (59.9%)

and External Test Set 2 shows a more balanced distribution (45.7% male, 54.3% female).

Regarding the eye side, the majority of cases involve both eyes (OS&OD), comprising 51.6% of the total for the internal datasets, while OS (left eye) and OD (right eye) have similar proportions. In External Test Set 1, all cases involve both eyes, while External Test Set 2 shows a smaller proportion of cases involving both eyes (52.9%). The report length shows some variation, with the internal datasets having an average report length of 99.6-99.7 words, while External Test Set 1 has a much shorter report length (28.6 words) and External Test Set 2 has a much longer average report length (132.8 words). The diagnostic distribution is also summarized, with the most common conditions across all datasets being retinal detachment (RD) (33.1% of total cases), followed by vitreous hemorrhage (VH) (24.0%), high myopia (HM) (18.5%), and cataract (10.7%). Other conditions like uveal melanoma (UM), refractive error (RE), and retinoblastoma (RB) are much less common.

Evaluation of Reports

Figure 2 presents examples of grounded AI-generated reports for real-world ophthalmic cases, including retinoblastoma, retinal detachment, cataract, and choroidal melanoma. In the left column, original ocular ultrasound images are displayed alongside the segmentation results from the VLS model, where the blue-filled areas represent the segmented lesions, and green dashed circles indicate the lesion annotations made by senior ophthalmologists. The second column features the written reports provided by ophthalmologists, and the third column shows the corresponding reports generated by the VLS system. These examples highlight the model's ability to generate clinically relevant reports that align closely with ophthalmologists' observations, demonstrating its potential for assisting in diagnostic and reporting tasks in real-world clinical settings.

The performance of report generation was evaluated in both the internal and external test sets, as shown in **Figure 3**. In the internal test set (**Figure 3 A-E**), both VLS and VL models achieved strong performance, with VL obtaining slightly higher BLEU4 (71.32 vs. 66.37) and ROUGE scores (84.49/75.99 vs.

82.39/73.49), indicating high fluency and accuracy in generating reports. These models processed samples at an efficient rate, with Lora_Qwen2.5vl achieving 0.16 samples per second and VL at 0.194 samples per second, significantly outperforming other models such as Qwen2.5vl and Llava_onevision, which processed at 0.043 and 0.055 samples per second, respectively.

In the external test sets, both models showed promising results (**Figure 3 F-J**). In External Test Set 1, VLS slightly outperformed VL (**Figure 3 F–I**). In External Test Set 1, VLS substantially outperformed VL across all NLG metrics, achieving a BLEU4 score of 85.36 compared to 64.47 for VL. Similarly, VLS achieved higher ROUGE-1 (88.45 vs. 75.75), ROUGE-2 (84.75 vs. 66.08), and ROUGE-L (90.37 vs. 75.76). Processing efficiency was comparable, with VL slightly faster (0.168 vs. 0.132 samples per second). In External Test Set 2, VLS again outperformed VL, though with a narrower margin: BLEU4 was 73.77 for VLS vs. 53.80 for VL, ROUGE-1 was 82.97 vs. 70.98, ROUGE-2 was 76.79 vs. 57.63, and ROUGE-L was 84.54 vs. 66.83. Processing speeds were again slightly higher for VLS (0.159 vs. 0.182 samples per second).

Evaluation of Segmentation

Figure 4 evaluates the zero-shot segmentation accuracy of the VLS model, prioritizing the dice coefficient as the primary metric, complemented by F1 score and specificity with 95% confidence intervals (CI) to better capture segmentation performance under potential class imbalance. Across various ophthalmic conditions, such as vitreous hemorrhage, cataract, uveal melanoma, retinoblastoma, retinal detachment, and high myopia, the model achieved a mean dice coefficient of 59.6% (95% CI: 51.2–68.0) and a mean F1 score of 59.6% (95% CI: 50.7–68.3) in the internal test set, with high specificity (99.3%) but variable sensitivity (63.9%). Performance was strongest for cataract (dice = 69.0%) and UM (dice = 68.2%).

When compared with Grounding DINO-US-SAM, the VLS model achieved a similar mean dice coefficient (59.6% vs. 57.7%), demonstrating broadly comparable overall segmentation accuracy. VLS showed substantial advantages for cataract (69.0% vs. 38.4%) and UM (68.2% vs. 52.5%), while DINO-US-SAM outperformed VLS in RD (77.4% vs. 53.3%) and HM (68.0% vs. 49.4%). Performance for VH was similar across models (62.3% vs. 64.6%). These results suggest that VLS and DINO-US-SAM offer complementary strengths across different ophthalmic conditions.

Figure 4 also provides the external evaluation of the VLS model's segmentation accuracy on two external test sets. While the model performs well across different conditions, the results show variability between the test sets. In External Test Set 1, the model demonstrates dice coefficient of 50.2%, the mean specificity is 97.8%. In External Test Set 2, the model's performance decreases slightly, with a mean Dice coefficient of 51.5%, and the mean specificity is 97.7%. Notably, conditions such as HM exhibit higher performance in external validation, especially in terms of F1 score.

Evaluation of Diagnostic Performance

Figure 5 summarizes the diagnostic performance of residents AI-aided modes for various ocular conditions across three test sets: internal test set, external test set 1, and external test set 2. The diagnostic accuracy (ACC), sensitivity, and specificity are presented with 95% CI for each condition. For cataracts, the diagnostic accuracy was high in the internal test set (93.6%), but sensitivity was low (66.47%), indicating that the model could miss some cases. In the external test set 1, the accuracy dropped to 71.88%, with a sensitivity of just 19.05%. For vitreous hemorrhage, the VLS model significantly improved diagnostic accuracy, especially in the external test sets, achieving 94.38% in external test set 1, though sensitivity remained low (25.0%). In contrast, specificity remained high (98.03%) in this case. In the case of high myopia, the performance in the internal test set was good, with an accuracy of 86.65% and high specificity (91.36%). However, the sensitivity was relatively lower (68.26%), especially in external test set 1, where sensitivity was 100% but specificity was only 18.59%. For uveal melanoma, diagnostic accuracy was near-perfect in the internal and external test sets, with 99.91% in the internal test set, but sensitivity was 0%, indicating a major issue in detecting this condition. Refractive error and retinoblastoma showed excellent specificity (100%) in the external test sets, but diagnostic accuracy varied across different test sets. For retinal detachment, diagnostic performance improved across all test sets, especially in external test set 2, where specificity reached 100%. Overall, VLS model diagnosis achieved 90.59% accuracy in the internal set, and 71.87% on the external datasets, though sensitivity and specificity varied notably across test sets.

Evaluation by Ophthalmologists

In Domain 1, which evaluated the extent of inappropriate content, the majority of reports generated by both AI systems and human physicians were rated as containing "None" inappropriate content (**Figure 6A**). Specifically, the SO produced the most appropriate reports (mean score: 2.70 in Chinese, 2.70 in English), followed by VLS (2.65 in Chinese, 2.68 in English) and VL (2.52 in Chinese, 2.61 in English). JO had the lowest appropriateness (2.39 in Chinese, 2.44 in English). Reports with "Present, little clinical significance" inappropriate content were more frequently observed in reports by the JO (38% in Chinese and 32% in English evaluations) compared to other systems, while reports with "Present, substantial clinical significance" inappropriate content were most commonly found in VL (highest risk profile, with inappropriate content score of 2.61 in English).

For Domain 2, which assessed the extent of missing content, the VLS AI system demonstrated superior performance (2.61 in Chinese, 2.17 in English), and the SO demonstrated consistently strong results (2.57 in Chinese, 2.63 in English) (**Figure 6A**). Interestingly, the JO showed better performance in English evaluation (2.40), compared to Chinese (2.34). Reports with "Present, little clinical significance" missing content were most frequently observed in reports by the JO (48% in Chinese evaluation), while reports with "Present, substantial clinical significance" missing content were most commonly found in VL (lowest score: 1.73 in English evaluation).

In Domain 3 (**Figure 6A**), the SO demonstrated the highest percentage of reports with low harm likelihood (2.82 in Chinese, 2.79 in English), followed by the VLS (2.76 in Chinese, 2.83 in English). Notably, VL had the lowest safety scores (1.73 in Chinese, 1.95 in English), indicating the greatest risk of harmful recommendations.

The total scores comparison (**Figure 6B**) revealed statistically significant differences between the AI systems and human physicians. In the Chinese language evaluation, the SO achieved the highest score of 8.09, followed closely by VLS (8.02), both significantly outperforming JO (7.13) and VL (6.64, $p < 0.001$). Similarly, in the English language evaluation, SO achieved the highest score of 8.12, followed by VLS with a score of 7.68, both of which significantly outperformed VL (6.29, $p < 0.001$) and JO (7.34).

Cost-effective Analysis

The total reading time (measured in minutes) showed substantial differences between AI systems and human physicians (**Figure 7**). The JO required the longest total reading time (approximately 163 minutes), followed by the SO (approximately 122 minutes). In contrast, the AI systems VLS and VL demonstrated significantly shorter total reading times of approximately 10.3 minutes and 8.7 minutes, respectively. This reflects a 15-18-fold reduction in reading time by AI systems compared to human physicians.

The per ophthalmic report reading time (measured in seconds) exhibited a similar pattern. Both JO and SO required substantially longer times per report (~98.0 and 70.4 seconds, respectively), as visualized in the violin plots showing their distributions concentrated in the higher range. The AI systems VLS and VL demonstrated markedly shorter per-report reading times (6.2 and 5.2 seconds, respectively), with their distributions concentrated in the lower range. The violin plot distributions also indicate greater consistency in reading times for AI systems compared to the wider variability observed in reports by human physicians.

The total cost analysis (measured in USD) revealed that the SO incurred the highest expenses (approximately $39.4), followed by the JO (approximately $28.3). In contrast, the AI systems VLS and VL demonstrated substantially lower total costs of approximately $1.3 and $1.0, respectively. This represents an approximately 30-40-fold difference in total cost between human physicians and AI systems. The per ophthalmic report cost followed a similar pattern. The AI systems VLS and VL demonstrated significantly lower per-report costs.

Human–AI Comparative Study

In total, 200 ophthalmic ultrasound cases encompassing seven diagnostic categories were independently evaluated by three junior ophthalmologists (JO1–JO3), both with and without AI assistance. When unaided, diagnostic accuracies were 88.0% for JO1, 84.0% for JO2, and 87.0% for JO3, corresponding to an average accuracy of 86.3%.

Following AI assistance, diagnostic performance improved markedly across all three participants. JO1 achieved 92.0% accuracy (+4%), while JO2 and JO3 both reached 96.0% accuracy, reflecting relative gains of +12% and +9%, respectively. The overall average accuracy with AI support was 94.7%, representing an absolute improvement of 8.3% compared to unaided performance.

At the disease level, AI assistance reduced misclassifications in common entities such as cataract, VH, and RD, while also stabilizing performance in less prevalent conditions, including RB and UM. Importantly, diagnostic sensitivity for high-morbidity conditions such as RD and VH improved consistently with AI support.

Discussion

The accurate and efficient interpretation of ophthalmic ultrasound images is crucial for early diagnosis and management of eye diseases. In this study, we have proposed an innovative approach that combines VLS with the SAM to generate grounded reports in ophthalmic ultrasound interpretation. For the interface of the VLS system and the demonstration of two analytical cases, please refer to Supplementary Movie 1. This approach leverages the strengths of advanced AI technologies to bridge the gap between visual data and clinical insights, offering a more reliable, scalable, and efficient solution for ophthalmic image analysis.

This study demonstrates the effectiveness of the VLS Model for ophthalmic ultrasound interpretation, which outperforms traditional methods, including reports generated by junior ophthalmologists and senior ophthalmologists. The VLS model achieved a BLEU4 score of 66.37 and a ROUGE-2 score of 73.49 in the internal test set, showing strong performance in both fluency and accuracy. More importantly, in the external test sets, the VLS model consistently outperformed the VL model, achieving markedly higher scores across all NLG metrics, including BLEU4 (85.36 vs. 64.47) and ROUGE-L (90.37 vs. 75.76) in External Test Set 1, and BLEU4 (73.77 vs. 53.80) and ROUGE-L (84.54 vs. 66.83) in External Test Set 2. These results underscore the model's strong generalizability across diverse datasets, despite variations in imaging devices and reporting practices. This highlights the potential of AI to automate ophthalmic report generation, offering superior efficiency and quality compared to expert-driven methods. Clinical effectiveness evaluations showed promising results, with the AI system achieving 94.38% accuracy for vitreous hemorrhage in external test set 1. However, sensitivity was low (25%), highlighting the need for further refinement to improve detection in certain conditions. For cataracts, the model reached 93.6% accuracy but showed reduced sensitivity, indicating challenges in detecting all cases. Overall, AI-aided diagnostic accuracy was 90.59%, a notable achievement in ophthalmic AI. Taken together, these findings suggest that the VLS model not only provides superior report fluency and content accuracy but also holds significant potential to improve efficiency and diagnostic consistency in real-world ophthalmic practice. Nonetheless, challenges remain in enhancing sensitivity for specific conditions and in addressing the variability introduced by heterogeneous imaging devices. Future work should therefore focus on multi-device domain adaptation, standardized reporting protocols, and larger-scale prospective validation to fully realize the clinical utility of AI-assisted ophthalmic ultrasound interpretation.

When compared to JO, the VLS system performed better in terms of the proportion of reports with no inappropriate content, with 74% of reports in Chinese and 79% in English evaluations showing no inappropriate content. The VLS model also excelled in detecting and reducing missing content, with 70% of reports in Chinese evaluations containing no missing information, further emphasizing its effectiveness. The VLS model reduced report costs by 40-fold compared to senior ophthalmologists, highlighting its practical value in resource-limited settings. This is especially valuable in ophthalmology, where expert interpretation is both time-consuming and expensive[20]. Additionally, the VLS model processed samples at a rate of 0.16 per second, demonstrating its efficiency in real-world clinical settings.

When compared to prior studies in the medical AI field, our approach offers a notable advancement by integrating VLM with SAM, an emerging methodology that enables precise image segmentation and enhanced understanding of medical terminology. BiomedGPT is an open-source, lightweight generalist AI model for biomedical tasks, achieving state-of-the-art performance in radiology question answering, report generation, and summarization, with potential to enhance diagnosis and workflow efficiency[21]. EchoCLIP is a vision-language model for echocardiography that effectively assesses cardiac function, identifies devices, and enables patient identification, advancing AI-driven preliminary interpretation of echocardiographic findings[17]. Previous works have explored AI models for ophthalmic image analysis, but few have integrated these two domains into a cohesive, closed-loop system that not only segments the images accurately but also generates clinically relevant and context-aware reports. Antaki et al. evaluated the performance of the Gemini Pro VLM for detecting macular diseases from OCT scans, showing limited feature detection capabilities but strong language consistency, highlighting the potential for VLMs in ophthalmology with further validation[22]. Chen et al. presented an AI-based framework for automated fundus fluorescein angiography interpretation, achieving strong performance with a BERTScore of 0.70 and F1 scores of 0.64-0.82 for detecting retinal conditions, alongside high accuracy and completeness in report generation as validated by ophthalmologists[23]. By leveraging both visual and linguistic models, our approach sets itself apart in the field of medical AI, offering a holistic solution that can bridge the gap between raw image data and clinical decision-making[24].

However, while the current study demonstrates promising results, several limitations must be addressed. Firstly, the dataset used in this study was somewhat limited in terms of the range of clinical conditions considered. While we observed strong performance in conditions like cataracts, vitreous hemorrhage, and retinal detachment, expanding the dataset to include a broader range of clinical diseases, such as diabetic retinopathy, glaucoma, and uveal melanoma, would enhance the model's robustness. Incorporating additional imaging modalities like optical coherence tomography (OCT) and fundus photography could further support a comprehensive multi-modal AI system[25]. This system could simultaneously analyze various types of imaging data, offering a unified platform for interpreting a broader range of ophthalmic diseases[26,27]. Although the VLS model demonstrated promising segmentation performance across various ophthalmic conditions, there is still room for improvement, particularly in achieving higher precision for certain conditions such as vitreous hemorrhage and retinal detachment. The model's performance showed variability between internal and external test sets, indicating that it may be sensitive to changes in dataset characteristics. This suggests that further refinement is needed to enhance its robustness and accuracy. Continued advancements in model architecture, training techniques, and data diversity will be essential for improving segmentation accuracy and ensuring that large models like VLS can achieve more reliable and precise results in clinical practice.

In addition, the system's performance can be negatively influenced by low-quality ultrasound images (e.g., blurred, noisy, or with poor contrast), which may reduce both segmentation accuracy and report reliability[28,29]. Another limitation lies in the detection of rare or small lesions, where limited training samples hinder the model's ability to generalize. These challenges highlight the need for future work to incorporate image quality assessment modules, targeted data augmentation, and the prospective inclusion of rare cases from multiple centers[30]. Such strategies would improve robustness and support the system's deployment in diverse real-world clinical settings. Moreover, although our model demonstrated faster processing times and reduced costs, integrating the system into real-world clinical workflows will require careful consideration of how to balance performance, speed, and resource use[20,31]. This could involve building AI systems that are not only efficient in terms of computational cost but also agile enough to adapt to different clinical settings with minimal latency[32,33].

In terms of future directions, we envision expanding our multi-modal AI system to support more complex clinical workflows by integrating data from various sources, such as patient history, and clinical notes. This would create a more holistic diagnostic tool capable of offering more accurate and comprehensive insights[34]. Our previous research shows that ChatGPT performs better with English prompts than Chinese prompts in diagnosing retinal vascular diseases, but still falls short of ophthalmologists, highlighting the need for further improvement in language models for clinical use[35]. Further, an important goal would be to create a truly agile AI system that can learn from new data in real-time, enabling continuous improvement as new clinical scenarios and imaging modalities emerge[36]. Furthermore, the importance of clinician-AI collaboration is evident, as demonstrated by the Flamingo-CXR AI system for automated chest radiograph report generation[37]. The system shows that AI-generated reports can be comparable to, or even preferable to, clinician reports in many cases, emphasizing the potential of human-AI teamwork to enhance report quality, reduce errors, and improve clinical workflows.

In conclusion, our study introduces a novel approach that integrates cutting-edge AI technologies to enhance the accuracy, efficiency, and cost-effectiveness of ophthalmic ultrasound interpretation. While the current model demonstrates significant promise, future work focusing on expanding the clinical dataset, integrating multi-modal imaging data, and improving the system's sensitivity for rare conditions will be essential for developing a more robust and comprehensive AI solution for ophthalmology. With continued advancements in AI technology, there is significant potential to revolutionize clinical practice by providing real-time, accurate, and cost-effective tools for healthcare professionals, ultimately improving patient outcomes in the field of ophthalmology.

Methods

This study developed a system combining a vision-language model for report generation and the SAM (Segmentation Anything Model) for lesion recognition (**Figure 1**). The methodology involved four key steps: (i) collecting and preparing a medical image dataset with lesion annotations for training; (ii) developing the VLM integrated with SAM for automatic lesion segmentation and report generation; (iii) conducting external validation to evaluate the system's performance, comparing it with baseline models; and (iv) performing clinical evaluations by ophthalmologists diagnosed and interpreted B-scan cases with the system's assistance, measuring diagnostic accuracy and reporting time, while six ophthalmologists assessed the clinical quality of the generated reports. This comprehensive evaluation ensured both technical performance and clinical applicability of the system.

Dataset Establishment

This study utilized data from three distinct datasets: SAHZU, FAHWM, and FAHZC, along with two external datasets, which includes a total of 63,979 images and 21,239 reports (**Table 1**). This study was performed in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committees of the Second Affiliated Hospital, Zhejiang University School of Medicine (No. Y2023-1073), the First Affiliated Hospital of Zhejiang Chinese Medical University No. 2024-KLS-583-02), and the First Affiliated Hospital of Wannan Medical College (No. Y2024-1015). The retrospective data were anonymized and approved by the Ethics Committees without the need for patient consent, while informed consent was obtained for the two external independent validation datasets. The datasets were divided into different subsets for training, validation, and testing. The SAHZU dataset consisted of 54,971 images and 12,649 reports, which were split into a training set (37,917 images), validation set (12,639 images), and test set (12,640 images). The FAHWM and FAHZC dataset included 652 images, and 163 images for the external test sets. The dataset establishment process adhered to rigorous inclusion criteria and underwent careful data handling to ensure data consistency and quality. For each dataset, images and reports were associated and reviewed by experts to ensure the highest accuracy in labeling and diagnosis.

The imaging machines used for data collection include the following models: For the SAHZU dataset, the imaging was performed using the Compact Touch, Cinescan, and Aviso machines from Quantel Medical (Clermont-Ferrand, France), as well as the ODM-2100 and MD2400S machines from MEDA Medical (Tianjin, China). The FAHWM dataset was acquired using the Compact Touch, Cinescan machines from Quantel Medical (Clermont-Ferrand, France). For the FAHZC dataset, the imaging was conducted with the SOVI Ophthalmic A/B Ultrasound Diagnostic Device SW-2100 from Tianjin SOVI. These machines were selected for their reliability and consistency in capturing high-quality ophthalmic ultrasound images, ensuring the validity and clinical relevance of the dataset.

The ophthalmic ultrasound images and corresponding free-text reports were retrieved from the picture archiving and communication system (PACS). Due to the use of different systems for external datasets, the image formats and resolutions varied. To ensure consistency, cases with poor image quality (e.g., significant blur or high noise caused by either pathological or technical issues) or those lacking corresponding free-text reports were excluded. After excluding these cases, the selected reports were initially reviewed by two residents (Z.L. and T.Y.), and then underwent quality control and final review by senior ophthalmologist (D.K. and K.J.) to correct any spelling errors and modify/exclude reports with incomplete information.

For all datasets, reports and images were annotated using a standardized annotation procedure. All annotations were performed by three ophthalmologists with over 5 years of experience, ensuring accuracy and consistency. The annotated reports were reviewed by two senior ophthalmologists with over 10 years of experience to further enhance the reliability of the dataset. Additionally, the lesion bounding boxes suggested by the SAM model were annotated by one senior ophthalmologist (D.K.) and reviewed by the other senior ophthalmologist (K.J.), ensuring the accuracy of lesion identification.

Model Development

In the development of the Vision-Language Model and SAM Integration, we employed two vision-language models: LLaVA One version and Qwen 2.5VL, both of which were trained to generate detailed and accurate medical reports from visual inputs. These models are designed to effectively handle complex medical imagery, making them well-suited for tasks involving nuanced and detailed descriptions such as those required in medical report generation. The LLaVA One version is specifically optimized for multi-modal tasks, combining image understanding and language generation capabilities[38]. It can effectively process medical images, extract relevant features, and generate coherent textual descriptions. Qwen 2.5VL, on the other hand, offers enhanced fine-tuning capabilities, enabling it to generate more contextually precise reports based on medical images[39]. This combination of models ensures high flexibility and accuracy in report generation across various clinical scenarios.

In addition to these models, we utilized LoRA (Low-Rank Adaptation) fine-tuning techniques to further enhance the performance of the vision-language models. LoRA enables efficient adaptation of pre-trained models by introducing low-rank decomposition layers, allowing for specialized fine-tuning with relatively fewer resources. This technique was particularly useful in tailoring the models to the specific task of medical report generation and lesion identification, without requiring extensive retraining from scratch.

As shown in Figure 1, lesion segmentation and report generation are tightly coupled within the proposed framework. First, during the model fine-tuning phase, we integrate bounding box information from images into existing medical reports. Using predefined prompts, multiple images, and curated medical reports, we fine-tune the Visual Language Model (VLM) to enable report generation and bounding box prediction capabilities. Then, during the inference stage, users input multiple images to obtain medical reports and potential lesion locations. If lesion locations exist, the original images are segmented using SAM to overlay the VLM-generated lesion bounding box information onto the images, yielding the final segmentation results. These bounding boxes serve a dual purpose: for the vision-language model (VLM, Qwen2.5VL), they provide visual context by describing lesion type, location, and clinical significance; for arbitrary segmentation models (SAM), they function as spatial prompts to guide the segmentation process.

The SAM model was not fine-tuned on ocular ultrasound images. Instead, its prompt encoder utilized bounding boxes to focus the segmentation, while its image encoder and mask decoder produced high-resolution lesion masks that refined boundaries beyond the initial box. By integrating these segmentation outputs into the VLM reasoning chain, the system ensured that textual reports were explicitly grounded in visual evidence. This linkage reduced the risk of hallucinations and enhanced clinical interpretability by providing explicit lesion localization within the report.

To further improve robustness, we implemented a rule-based arbitration mechanism when discrepancies arose between SAM and VLM outputs (e.g., the VLM narrative suggested "retinal detachment" while SAM highlighted a different region). The system jointly evaluated the segmentation confidence score (Dice-based) and the VLM confidence score (logit-based). If segmentation confidence was low, the VLM narrative was prioritized. If segmentation confidence was high but conflicted with the VLM interpretation, the final report explicitly noted the inconsistency (e.g., "segmentation suggests X, but textual interpretation indicates Y"). This design ensured that neither model operated in isolation and that potential conflicts were transparently reported rather than suppressed.

All the models were trained using two Nvidia V100 GPUs on the backend framework of PyTorch, leveraging distributed parallelism to accelerate the training process. The Adam optimizer was employed for optimization, with initial learning rates set at 5e-5 for the visual extractor and 1e-4 for all other model parameters. The learning rate was decayed by a factor of 0.8 at the end of each epoch to facilitate stable convergence. For the diagnosis-supervised contrastive loss, the weight α was set to 0.2, balancing the contribution of the contrastive loss with other components of the total loss function. This training configuration ensured efficient model optimization while maintaining high performance across both vision-language processing and lesion segmentation tasks.

Automatic Evaluation of Reports

The performance of the four models described earlier was automatically evaluated using NLG (Natural Language Generation) and CE (Classification Evaluation) metrics on the SAHZU test set as well as two external test datasets. The NLG metrics utilized in this evaluation included Bilingual Evaluation Understudy 1 (BLEU), Metric for Evaluation of Translation with Explicit Ordering (METEOR), and ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation - Longest common subsequence). BLEU is commonly used to assess the quality of machine translation by measuring the overlap of n-grams (sequences of n words) between the generated text and the reference text. In this study, we computed BLEU1, BLEU2, BLEU3, and BLEU4 as part of our evaluation. METEOR builds upon BLEU by incorporating synonyms and paraphrases, thus offering a more flexible approach to evaluating the generated ultrasound reports. ROUGE-L, which emphasizes recall, is especially valuable for evaluating how well the generated text captures the essential ideas and key clinical information, making it ideal for assessing the coherence of complex medical descriptions.

Automatic Evaluation of Segmentation

The VLS model was evaluated for zero-shot segmentation accuracy across multiple ophthalmic conditions, including vitreous hemorrhage, cataract, uveal melanoma, retinoblastoma, retinal detachment, and high myopia. Model performance was assessed using metrics such as dice coefficient, sensitivity, specificity, and F1 score. The internal evaluation was conducted on a set of annotated images, while external evaluations were performed on two separate test sets to examine the model's generalizability. Statistical analysis included the calculation of 95% confidence intervals for all metrics, allowing for a comprehensive assessment of the model's accuracy and reliability in both internal and external settings.

For benchmarking, we additionally compared the VLS framework with the Grounding DINO-US-SAM model[40]. Specifically, we employed the publicly available implementation of Grounding DINO coupled with SAM for zero-shot segmentation. The same annotated internal dataset was used for evaluation to ensure consistency across models. Both VLS and DINO-US-SAM were tested under identical experimental conditions, including preprocessing steps, evaluation metrics, and confidence interval estimation. This setup enabled a direct and reproducible comparison of segmentation performance between the two approaches, highlighting complementary strengths across different ophthalmic conditions.

Evaluation of diagnostic accuracy

We developed an online tool for drafting and diagnosing ophthalmic ultrasound reports. This tool simulates real-world PACS viewer functionalities, including image switching, zooming, labeling, measurement adjustments, and contrast modifications, without displaying any patient-specific information. In the AI-assisted mode, an AI-generated report is displayed when viewers examine the corresponding ultrasound images. Viewers have the option to either adopt, modify, or discard the AI-generated report based on their own clinical observations and expertise. In the standard template-aided mode, preformatted reports corresponding to the seven ophthalmic conditions are presented as references. Viewers can then select an appropriate diagnosis based on the displayed information. Additionally, the tool records both the time taken for diagnosis and report generation.

Human Evaluation of the Head-to-Head Comparison

To assess the performance of the VLS compared to VL, junior ophthalmologists (JO), and senior ophthalmologists (SO) in providing management recommendations for ophthalmic cases, we curated a dataset comprising 100 clinical cases. These cases were randomly selected and used to evaluate the models' capabilities in both English and Chinese languages. In this evaluation, three core criteria were considered: the extent of inappropriate content, the extent of missing content, and the likelihood of potential harm in the management recommendations. The evaluations were performed by a panel of expert evaluators who rated the cases based on these criteria. The evaluators included six licensed ophthalmologists in total: three JOs and three SOs.

The ophthalmologists involved in this evaluation all hold valid medical licenses. The three JOs, each with 3-5 years of clinical experience, are familiar with a wide range of ophthalmic conditions and management strategies. The three SOs, with over 10 years of clinical experience, possess deep expertise in complex and rare ophthalmic conditions. Their extensive experience enables them to make well-rounded, informed clinical decisions, especially in complicated cases. This diverse panel of both junior and senior ophthalmologists, along with the AI models, provided valuable insights into the effectiveness of the VLS compared to human expertise in ophthalmology.

For each of the 100 cases, management recommendations were generated using the VLS, VL, and the two ophthalmologist groups (JO and SO). These recommendations were anonymized and then assessed by a separate panel of two expert ophthalmologists. The evaluations were carried out in both English and Chinese, with each language having a distinct expert panel. For the English evaluation, the panel consisted of bilingual ophthalmologists who rated the management recommendations based on the pre-established criteria. The assessment employed a box plot to display the total scores, which incorporated the extent of inappropriate content, missing content, and potential harm. Statistical comparisons between the four groups (VLS, VL, JO, and SO) were performed using two-sided Friedman tests. Post-hoc pairwise comparisons were carried out using two-sided Wilcoxon signed-rank tests, with P-values adjusted for multiple comparisons using the Bonferroni method.

Additionally, to ensure robust and reliable evaluations, a separate ablation study was conducted to compare the performance of VLS and VL in both languages. For detailed evaluation methods and criteria, please refer to Supplementary Table 1-3.

Cost-Effectiveness Analysis

A cost-effectiveness analysis was conducted to compare the total reading time and the total cost for each AI system and human physician. Total reading time was measured in minutes, and per-report reading time was recorded in seconds. Each system's performance was analyzed in terms of how quickly reports could be processed, with a focus on the difference between human physicians and AI systems. Additionally, the total cost incurred by each system was evaluated. The cost was calculated based on the time and resources required to process the reports. The analysis allowed for a comparison of the economic efficiency of the AI systems versus human physicians, with particular attention to the differences in both total costs and per-report costs.

Human–AI Comparative Study Design

To evaluate the impact of AI-assisted interpretation on diagnostic performance, we conducted a human–AI comparative experiment. Three ophthalmology residents independently reviewed and interpreted 200 ophthalmic ultrasound cases (Supplementary Table 4). For each case, the residents generated a structured diagnostic report and provided a final diagnosis without access to the VLS AI system. After a washout period, the same residents re-evaluated the identical set of cases with the assistance of the AI, which automatically generated preliminary diagnostic suggestions and structured outputs. The final diagnoses made by each resident, both with and without AI support, were compared against the gold standard established by two senior ophthalmologists. Diagnostic accuracy rates were calculated for each resident in the unassisted and AI-assisted conditions, enabling quantitative assessment of the effect of AI on diagnostic performance.

Statistical Analysis

Clinical efficacy was assessed by diagnostic accuracy, sensitivity, specificity, and report-writing time across all diseases and within disease subgroups. Confidence intervals (CIs) for these metrics were calculated using the Wilson score interval method. To compare the performance of AI systems and human physicians, descriptive statistics (mean, standard deviation, and percentages) were used to summarize the total scores, inappropriate content, missing content, and harm likelihood. For comparisons of total scores, independent t-tests or ANOVA were used, depending on the data distribution. Post-hoc pairwise comparisons with Bonferroni correction were applied where necessary. Chi-square tests were used to assess significant differences in the distribution of content and harm categories between systems. For cost and time-related outcomes, independent t-tests were employed to compare total reading time and costs between AI systems and human physicians. Per-report costs and reading times were also compared using appropriate statistical tests. All statistical tests were two-sided, with p-values < 0.05 considered significant. Bonferroni correction was applied for multiple comparisons. Statistical analyses were performed using R software (version 4.2.1), and results are presented as mean ± standard deviation for continuous variables and percentages for categorical data.

The datasets used and analyzed in this study are not publicly available due to patient privacy considerations, but they can be obtained from the corresponding author upon reasonable request and with appropriate institutional approvals. The source code used in this study has been made publicly available at: https://github.com/Qix-Sun/Vit.

Declaration statements

Data availability

The datasets used and analyzed in this study are available from the corresponding author upon reasonable request.

Code availability

The code is available at https://github.com/Qix-Sun/Vit.

Acknowledgements

This study was supported by National Natural Science Foundation of China (82201195).

Author contribution

Kai Jin conceived the study, designed the overall framework, and secured funding. Qixuan Sun was responsible for the algorithm, model development, and statistical analysis. Daohuan Kang performed the statistical analysis and contributed to the interpretation of the results. Daohuan Kang, Wenzheng Han and Yi Zhang conducted the clinical validation. Ziyao Luo, Tao Yu, Wenzheng Han and Yi Zhang collected and organized the data. Kai Jin wrote and revised the initial manuscript. Meng Wang, Danli Shi and Andrzej Grzybowski provided feedback and suggestions for manuscript revisions.

Competing interest

None to declare.

Reference

1 Rao, V. M. *et al.* Multimodal generative AI for medical image interpretation. *Nature* **639**, 888-896, doi:10.1038/s41586-025-08675-y (2025).

2 Propst, S. L. *et al.* Ocular Point-of-Care Ultrasonography to Diagnose Posterior Chamber Abnormalities: A Systematic Review and Meta-analysis. *JAMA network open* **3**, e1921460-e1921460, doi:10.1001/jamanetworkopen.2019.21460 (2020).

3 Maheshwari, A. & Finger, P. T. Cancers of the eye. *Cancer metastasis reviews* **37**, 677-690, doi:10.1007/s10555-018-9762-9 (2018).

4 Resnikoff, S., Felch, W., Gauthier, T. M. & Spivey, B. The number of ophthalmologists in practice and training worldwide: a growing gap despite more than 200,000 practitioners. *The British journal of ophthalmology* **96**, 783-787, doi:10.1136/bjophthalmol-2011-301378 (2012).

5 Resnikoff, S. *et al.* Estimated number of ophthalmologists worldwide (International Council of Ophthalmology update): will we meet the needs? *The British journal of ophthalmology* **104**, 588-592, doi:10.1136/bjophthalmol-2019-314336 (2020).

6 Abràmoff, M. D. *et al.* Foundational Considerations for Artificial Intelligence Using Ophthalmic Images. *Ophthalmology* **129**, e14-e32, doi:10.1016/j.ophtha.2021.08.023 (2022).

7 Li, Z. *et al.* Artificial intelligence in ophthalmology: The path to the real-world clinic. *Cell reports. Medicine* **4**, 101095, doi:10.1016/j.xcrm.2023.101095 (2023).

8 Jin, K. *et al.* Integration of smartphone technology and artificial intelligence for advanced ophthalmic care: A systematic review. *Advances in Ophthalmology Practice and Research* **4**, 120-127, doi:https://doi.org/10.1016/j.aopr.2024.03.003 (2024).

9 Mihalache, A. *et al.* Accuracy of an Artificial Intelligence Chatbot's Interpretation of Clinical Ophthalmic Images. *JAMA ophthalmology* **142**, 321-326, doi:10.1001/jamaophthalmol.2024.0017 (2024).

10 Yu, T. *et al.* A Systematic Review of Advances in AI-Assisted Analysis of Fundus Fluorescein Angiography (FFA) Images: From Detection to Report Generation. *Ophthalmology and therapy* **14**, 599-619, doi:10.1007/s40123-025-01109-y (2025).

11 Shandhi, M. M. H. & Dunn, J. P. AI in medicine: Where are we now and where are we going? *Cell reports. Medicine* **3**, 100861, doi:10.1016/j.xcrm.2022.100861 (2022).

12 Wang, Y. *et al.* Automated classification of multiple ophthalmic diseases using ultrasound images by deep learning. *British Journal of Ophthalmology* **108**, 999-1004, doi:10.1136/bjo-2022-322953 (2024).

13 Ye, X. *et al.* Ocular Disease Detection with Deep Learning (Fine-Grained Image Categorization) Applied to Ocular B-Scan Ultrasound Images. *Ophthalmology and therapy* **13**, 2645-2659, doi:10.1007/s40123-024-01009-7 (2024).

14 Gao, Z. *et al.* Automatic interpretation and clinical evaluation for fundus fluorescein angiography images of diabetic retinopathy patients by deep learning. *British Journal of Ophthalmology* **107**, 1852-1858 (2023).

15 Lim, G., Elangovan, K. & Jin, L. Vision language models in ophthalmology. *Current opinion in ophthalmology* **35**, 487-493, doi:10.1097/icu.0000000000001089 (2024).

16 Li, J. *et al.* Integrated image-based deep learning and language models for primary diabetes care. *Nature medicine* **30**, 2886-2896, doi:10.1038/s41591-024-03139-8 (2024).

17 Christensen, M., Vukadinovic, M., Yuan, N. & Ouyang, D. Vision–language foundation model for echocardiogram interpretation. *Nature medicine* **30**, 1481-1488, doi:10.1038/s41591-024-02959-y (2024).

18 Ma, J. *et al.* Segment anything in medical images. *Nature communications* **15**, 654, doi:10.1038/s41467-024-44824-z (2024).

19 Wu, J. *et al.* Medical SAM adapter: Adapting segment anything model for medical image segmentation. *Medical image analysis* **102**, 103547, doi:10.1016/j.media.2025.103547 (2025).

20 Yan, Y. *et al.* Clinical evaluation of deep learning systems for assisting in the diagnosis of the epiretinal membrane grade in general ophthalmologists. *Eye* **38**, 730-736, doi:10.1038/s41433-023-02765-9 (2024).

21 Zhang, K. *et al.* A generalist vision–language foundation model for diverse biomedical tasks. *Nature medicine* **30**, 3129-3141, doi:10.1038/s41591-024-03185-2 (2024).

22 Antaki, F., Chopra, R. & Keane, P. A. Vision-Language Models for Feature Detection of Macular Diseases on Optical Coherence Tomography. *JAMA ophthalmology* **142**, 573-576, doi:10.1001/jamaophthalmol.2024.1165 (2024).

23 Chen, X. *et al.* FFA-GPT: an automated pipeline for fundus fluorescein angiography interpretation and question-answer. *NPJ Digit Med* **7**, 111, doi:10.1038/s41746-024-01101-z (2024).

24 Lu, Y. & Wang, A. Integrating language into medical visual recognition and reasoning: A survey. *Medical image analysis* **102**, 103514, doi:10.1016/j.media.2025.103514 (2025).

25 Jin, K., Yu, T. & Grzybowski, A. Multimodal artificial intelligence in ophthalmology: Applications, challenges, and future directions. *Survey of ophthalmology*, doi:10.1016/j.survophthal.2025.07.003.

26 Jin, K. *et al.* Multimodal deep learning with feature level fusion for identification of choroidal neovascularization activity in age-related macular degeneration. *Acta ophthalmologica*, doi:10.1111/aos.14928.

27 Liu, J. *et al.* Challenges in AI-driven Biomedical Multimodal Data Fusion and Analysis. *Genomics, proteomics & bioinformatics*, doi:10.1093/gpbjnl/qzaf011 (2025).

28 Jin, K. *et al.* MSHF: A Multi-Source Heterogeneous Fundus (MSHF) Dataset for Image Quality Assessment. *Scientific data* **10**, 286, doi:10.1038/s41597-023-02188-x (2023).

29 Wu, H. *et al.* Diabetic Retinopathy Assessment through Multitask Learning Approach on Heterogeneous Fundus Image Datasets. *Ophthalmology science* **5**, doi:10.1016/j.xops.2025.100755 (2025).

30 Grzybowski, A., Jin, K. & Wu, H. Challenges of artificial intelligence in medicine and dermatology. *Clinics in dermatology* **42**, 210-215, doi:https://doi.org/10.1016/j.clindermatol.2023.12.013 (2024).

31 Wenderott, K., Krups, J., Zaruchas, F. & Weigl, M. Effects of artificial intelligence implementation on efficiency in medical imaging—a systematic literature review and meta-analysis. *NPJ Digit Med* **7**, 265, doi:10.1038/s41746-024-01248-9 (2024).

32 Ciecierski-Holmes, T., Singh, R., Axt, M., Brenner, S. & Barteit, S. Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: a systematic scoping review. *NPJ Digit Med* **5**, 162, doi:10.1038/s41746-022-00700-y (2022).

33 Wu, H., Jin, K., Yip, C. C., Koh, V. & Ye, J. A systematic review of economic evaluation of artificial intelligence-based screening for eye diseases: From possibility to reality. *Survey of ophthalmology* **69**, 499-507, doi:10.1016/j.survophthal.2024.03.008 (2024).

34 Xu, K. *et al.* Digital twins in ophthalmology: Concepts, applications, and challenges. *Asia-Pacific journal of ophthalmology (Philadelphia, Pa.)*, 100205, doi:10.1016/j.apjo.2025.100205 (2025).

35 Liu, X. *et al.* Uncovering language disparity of ChatGPT on retinal vascular disease classification: cross-sectional study. *Journal of medical Internet research* **26**, e51926 (2024).

36 Ying, Z. *et al.* Real-Time AI-Assisted Insulin Titration System for Glucose Control in Patients With Type 2 Diabetes: A Randomized Clinical Trial. *JAMA network open* **8**, e258910, doi:10.1001/jamanetworkopen.2025.8910 (2025).

37 Tanno, R. *et al.* Collaboration between clinicians and vision–language models in radiology report generation. *Nature medicine* **31**, 599-608, doi:10.1038/s41591-024-03302-1 (2025).

38 Li, B. *et al.* Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326* (2024).

39 Bai, S. *et al.* Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923* (2025).

40 Rasaee, H., Koleilat, T. & Rivaz, H. Grounding DINO-US-SAM: Text-Prompted Multiorgan Segmentation in Ultrasound With LoRA-Tuned Vision–Language Models. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* **72**, 1414-1425, doi:10.1109/TUFFC.2025.3605285 (2025).

Figures

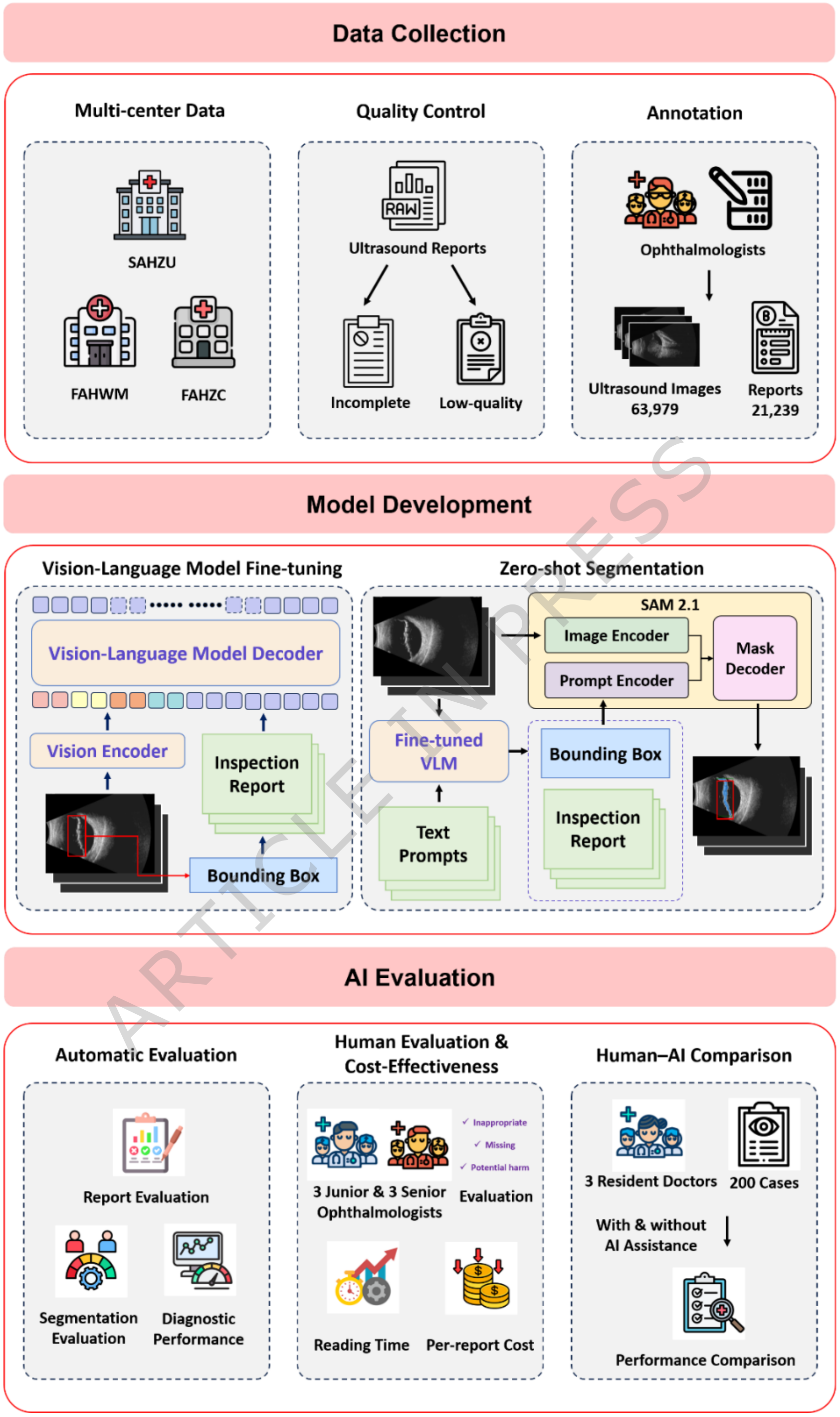


Figure 1. Overview of the study design. The Second Affiliated Hospital of Zhejiang University (SAHZU), the First Affiliated Hospital of Zhejiang Chinese Medical University (FAHZC), and the First Affiliated Hospital of Wannan Medical College (FAHWM);VLM: Vision-Language Models; Segment Anything

Model (SAM). The icons were created by Freepik, and permission for their use has been granted.

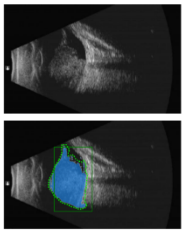
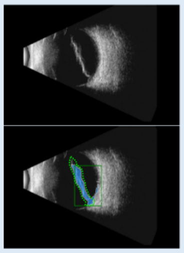
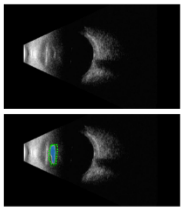
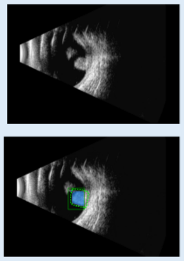
The segmented image	Clinician's report	AI report
	**Chinese:** 诊断结果为：视网膜母细胞瘤。检查结果为：左眼玻璃体少量光点光条左眼后脱离光带无，后运动无左眼后极附近球壁探及迂曲强回声光带隆起，后连视乳头左眼后极及下方玻璃体内探及实质性占位，内部回声欠均匀，隐约探及强回声光斑，后方伴声影。最后提示为：左眼玻璃体混浊左眼视网膜脱离左眼球内实质性占位，建议进一步检查。	**Chinese:** 诊断结果为：视网膜母细胞瘤。检查结果为：左眼玻璃体大量光点、光条左眼玻璃体内探及条索状光带与后极部球壁回声粘连左眼后极部球壁探及缝隙状隆起。最后提示为：左眼玻璃体混浊（左眼混浊性质请结合临床）左眼玻璃体视网膜粘连性改变左眼后极部球壁局限性隆起，请结合临床。
	English: The diagnosis was: retinoblastoma. Examination results were: left eye vitreous small amount of light spot light strip left eye after detachment light band no, posterior motion no left eye posterior pole near the bulge of the bulb wall detected tortuous strong echo light band, posterior connection to the optic papillae left eye posterior pole and below the intravitreous detected substantial occupations, the internal echogenicity of non-uniformity, vaguely detected strong echogenic light spot, the posterior accompanied by the acoustic shadows. The final suggestion was: left vitreous clouding, left retinal detachment, left intraocular space, further examination was recommended.	**English:** The diagnosis was: retinoblastoma. Examination findings were: large number of spots and streaks in the left vitreous body, streaks detected in the left vitreous body and echogenic adhesion to the posterior pole wall, and a slit-like bulge detected in the posterior pole wall of the left eye. The last suggestion was: left eye vitreous clouding (please combine the nature of clouding in the left eye with the clinic) left eye vitreoretinal adhesive changes left eye limited bulging of the posterior pole bulb wall of the left eye, please combine the clinic with the clinic.
	**Chinese:** 诊渐结果为：视网膜脱离。检查结果为：右眼玻璃体少量光点光条，右眼后脱离光带无，后运动无，右眼颞侧，上方，下方探及迂曲强回声光带隆起，后连视乳头。最后提示为：右眼玻璃体混浊，右眼视网膜脱离可能。	**Chinese:** 诊断结果为：视网膜脱离。检查结果为：右眼玻璃体少量光点、光条右眼后脱离光带无，最后提示为：右眼玻璃体混浊，右眼视网膜脱离可能。
	English: The diagnostic findings were: retinal detachment. The examination results were: a small amount of vitreous light spots and light stripes in the right eye, no posterior detachment light band in the right eye, no posterior motion, the right eye temporal, superior, and inferior to the detached eye detected a tortuous echogenic light band bulge, the posterior connection to the optic papilla. The final suggestion was: clouding of the right vitreous and possible retinal detachment of the right eye.	**English:** The diagnosis was: retinal detachment. The examination results were: a small number of light spots in the vitreous of the right eye, light strips in the right eye after detachment of the light band was absent, and the final suggestion was: vitreous clouding in the right eye and possible retinal detachment of the right eye.
	**Chinese:** 诊渐结果为：白内障。检查结果为：右眼玻璃体少量光点、光条，右眼后脱离光带细，后运动明显，视网膜未探及明显脱离光带。最后提示为：右眼玻璃体混浊伴右眼后脱离。	**Chinese:** 诊断结果为：白内障。检查结果为：右眼上方隐约探及光斑，右眼玻璃体轻度光点、光团可见，水晶体隐约隆起，右眼视网膜未探及明显脱离光带。最后提示为：右眼玻璃体少光点及光条。
	English: The diagnosis was: cataract. The examination results were: a small number of light spots and light strips in the vitreous of the right eye, thin light bands in the right eye after detachment, obvious posterior motion, and no obvious detachment bands were detected in the retina. The final suggestion was: vitreous clouding with posterior detachment in the right eye.	**English:** The diagnosis was: cataract. The examination findings were: light spots were vaguely detected above the right eye, light spots and light clusters were visible in the right vitreous, the crystal was vaguely elevated, and no obvious detached light bands were detected in the right retina. The final indication was: right eye vitreous body less light spots and light stripes.
	**Chinese:** 诊渐结果为：脉络膜黑色素瘤。检查结果为：右眼玻璃体少量光点、光条，右眼后脱离光带无，后运动无，视网膜未探及明显脱离光带，右眼鼻侧周边部球壁探及实质性隆起，边界清，内为较均匀中强回声。最后提示为：右眼玻璃体混浊，右眼球内实质性占位。	**Chinese:** 诊断结果为：脉络膜黑色素瘤。检查结果为：右眼玻璃体大量光点、光线条膜状光带与上方巩膜接触，右眼后脱离光带细，后运动无，视网膜未探及明显脱离光带。最后提示为：右眼玻璃体混浊，右眼球内实质性占位。
	English: The diagnostic findings were: choroidal melanoma. The examination results were: a small number of light spots and light strips in the vitreous body of the right eye, no posterior detachment of light bands in the right eye, no posterior motion, no obvious detachment of light bands detected in the retina, and a substantial bulge was detected in the bulbous wall of the right eye in the nasal periphery, with clear borders and relatively uniform medium-strong echoes. The final suggestion was: vitreous clouding in the right eye, and substantial intraocular space in the right eye.	**English:** Diagnostic findings were: choroidal melanoma. Examination results were: right eye vitreous large number of light spots, light strips of membranous light bands in contact with the upper sclera, right eye after the detachment of light bands thin, no posterior motion, the retina did not detect obvious detachment of light bands. The final suggestion was: vitreous clouding with retinal detachment in the right eye, and substantial intraocular space in the right eye.

Figure 2. Examples of grounded AI-generated reports for real-world ophthalmic cases, including retinoblastoma, retinal detachment, cataract, and choroidal melanoma. The left column displays original ocular ultrasound images with VLS segmentation results (blue-filled area) and lesion annotations made by ophthalmologists (green dashed circles). The second column contains the ophthalmologist's written report, while the third column shows the report generated by the VLS system.

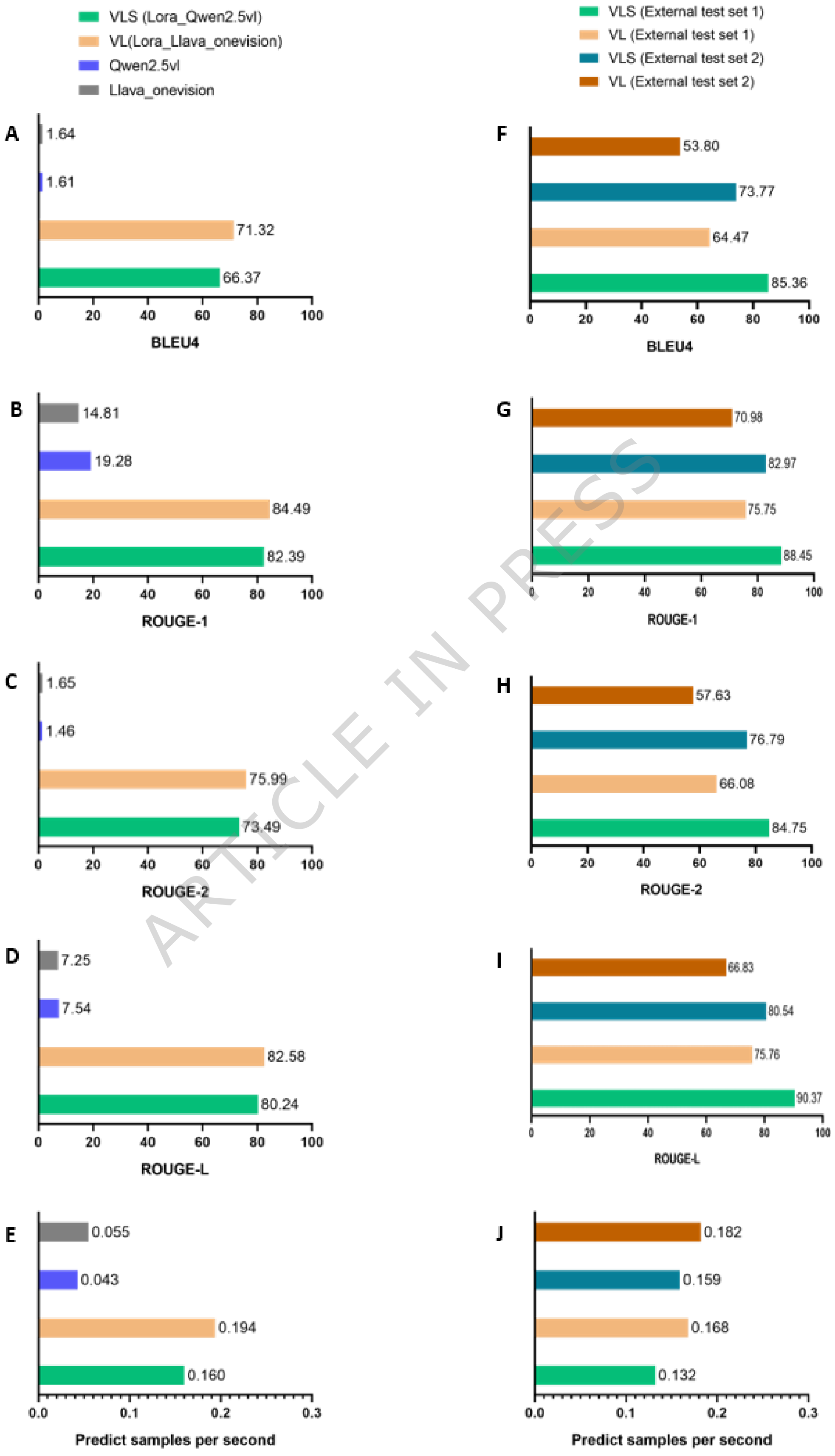

VLS (Lora_Qwen2.5vl)
VL(Lora_Llava_onevision)
Qwen2.5vl
Llava_onevision
VLS (External test set 1)
VL (External test set 1)
VLS (External test set 2)
VL (External test set 2)
A
1.64
1.61
71.32
66.37
0
20
40
60
80
100
BLEU4
F
53.80
73.77
64.47
85.36
0
20
40
60
80
100
BLEU4
B
14.81
19.28
84.49
82.39
0
20
40
60
80
100
ROUGE-1
G
70.98
82.97
75.75
88.45
0
20
40
60
80
100
ROUGE-1
C
1.65
1.46
75.99
73.49
0
20
40
60
80
100
ROUGE-2
H
57.63
76.79
66.08
84.75
0
20
40
60
80
100
ROUGE-2
D
7.25
7.54
82.58
80.24
0
20
40
60
80
100
ROUGE-L
I
66.83
80.54
75.76
90.37
0
20
40
60
80
100
ROUGE-L
E
0.055
0.043
0.194
0.160
0.0
0.1
0.2
0.3
Predict samples per second
J
0.182
0.159
0.168
0.132
0.0
0.1
0.2
0.3
Predict samples per second

Figure 3. Report generation performance in internal test set (A-E) and in external test sets (F-J). VLS = Vision-Language Segmentation model, VL = Vision-Language model.

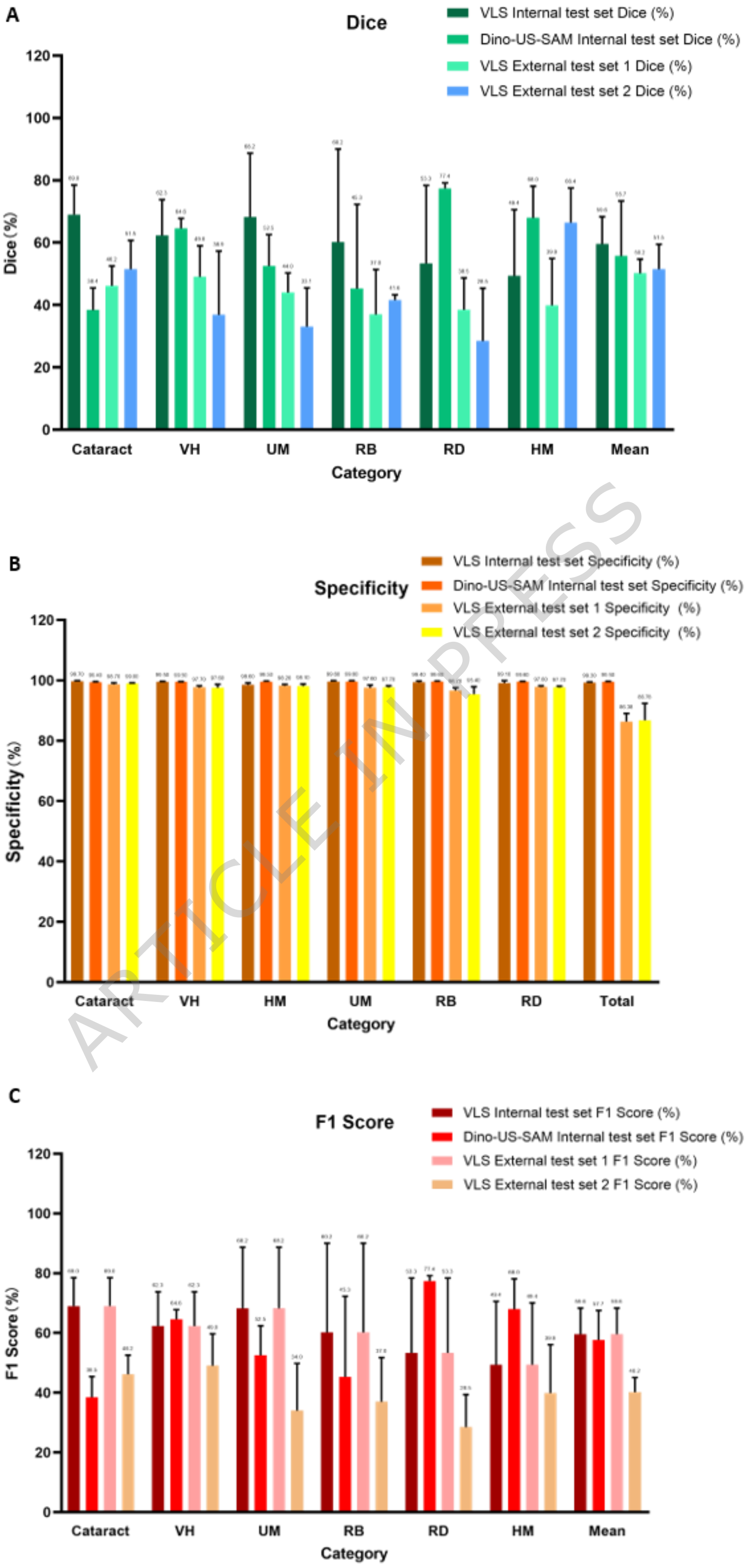

A
Dice
🟩 VLS Internal test set Dice (%)
🟩 Dino-US-SAM Internal test set Dice (%)
🟩 VLS External test set 1 Dice (%)
🟦 VLS External test set 2 Dice (%)
Dice (%)
0
20
40
60
80
100
120
Cataract
VH
UM
RB
RD
HM
Mean
Category
B
Specificity
🟫 VLS Internal test set Specificity (%)
🟧 Dino-US-SAM Internal test set Specificity (%)
🟧 VLS External test set 1 Specificity (%)
🟨 VLS External test set 2 Specificity (%)
Specificity (%)
0
20
40
60
80
100
120
Cataract
VH
HM
UM
RB
RD
Total
Category
C
F1 Score
🟥 VLS Internal test set F1 Score (%)
🟥 Dino-US-SAM Internal test set F1 Score (%)
🟥 VLS External test set 1 F1 Score (%)
🟧 VLS External test set 2 F1 Score (%)
F1 Score (%)
0
20
40
60
80
100
120
Cataract
VH
UM
RB
RD
HM
Mean
Category

Figure 4. Evaluation of segmentation accuracy for Vision-Language Segmentation (VLS) models and Dino-US-SAM models. VH = vitreous hemorrhage, HM = high myopia, UM = uveal melanoma, RE= refractive error, RB = retinoblastoma, RD = retinal detachment.

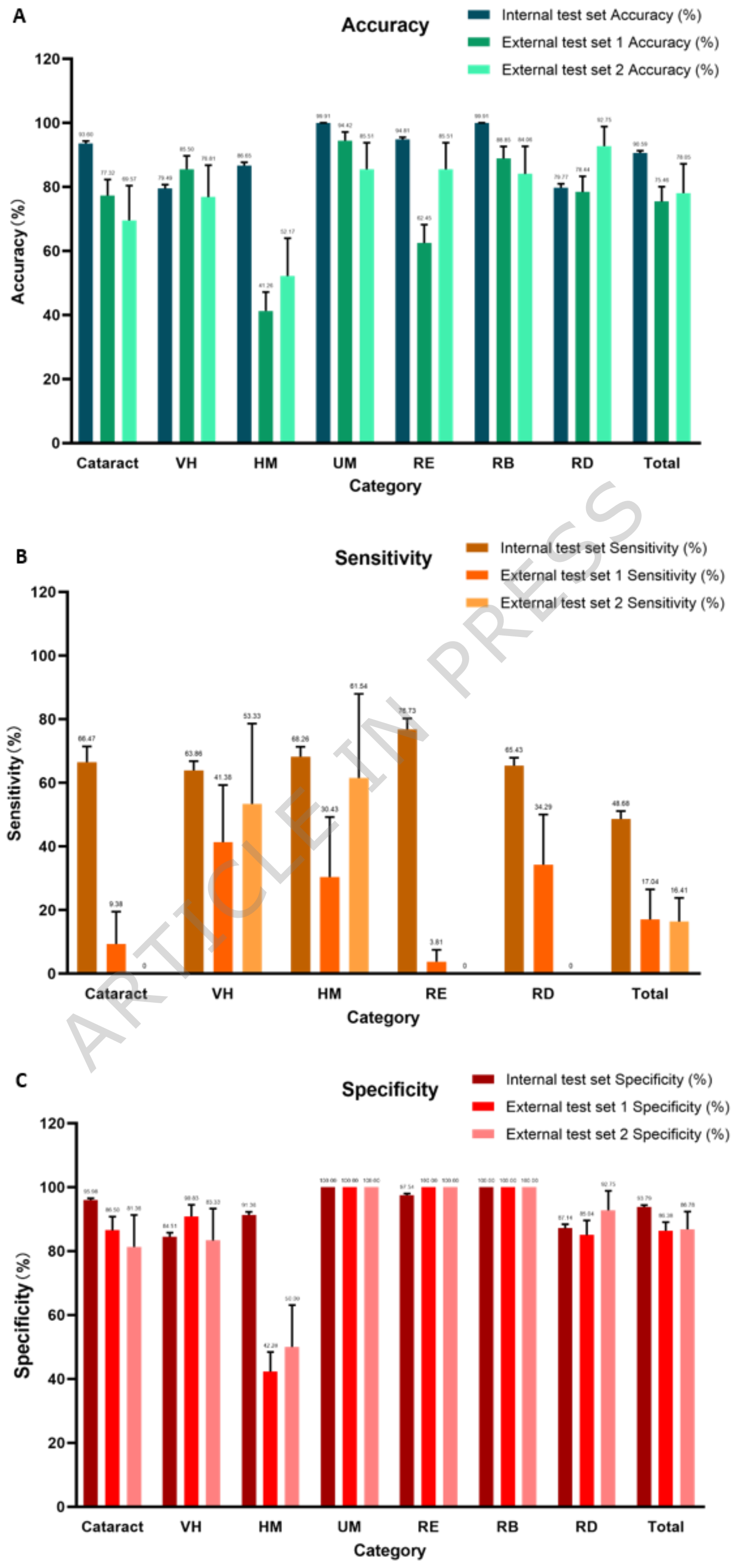

A
Accuracy
Internal test set Accuracy (%)
External test set 1 Accuracy (%)
External test set 2 Accuracy (%)
Accuracy (%)
0
20
40
60
80
100
120
93.60
77.32
69.57
79.49
85.50
76.81
86.65
41.26
52.17
99.91
94.42
85.51
94.81
62.45
85.51
99.91
88.85
84.06
79.77
78.44
92.75
90.59
75.46
78.05
Cataract
VH
HM
UM
RE
RB
RD
Total
Category
B
Sensitivity
Internal test set Sensitivity (%)
External test set 1 Sensitivity (%)
External test set 2 Sensitivity (%)
Sensitivity (%)
0
20
40
60
80
100
120
66.47
9.38
0
63.86
41.38
53.33
68.26
30.43
61.54
76.73
3.81
0
65.43
34.29
0
48.68
17.04
16.41
Cataract
VH
HM
RE
RD
Total
Category
C
Specificity
Internal test set Specificity (%)
External test set 1 Specificity (%)
External test set 2 Specificity (%)
Specificity (%)
0
20
40
60
80
100
120
95.98
86.50
81.38
84.51
90.83
83.33
91.38
42.28
50.00
100.00
100.00
100.00
97.54
100.00
100.00
100.00
100.00
100.00
87.14
85.04
92.75
93.79
86.38
86.78
Cataract
VH
HM
UM
RE
RB
RD
Total
Category

Figure 5. The diagnostic performance of Vision-Language Segmentation (VLS) models. VH = vitreous hemorrhage, HM = high myopia, UM = uveal melanoma, RE= refractive error, RB = retinoblastoma, RD = retinal detachment.

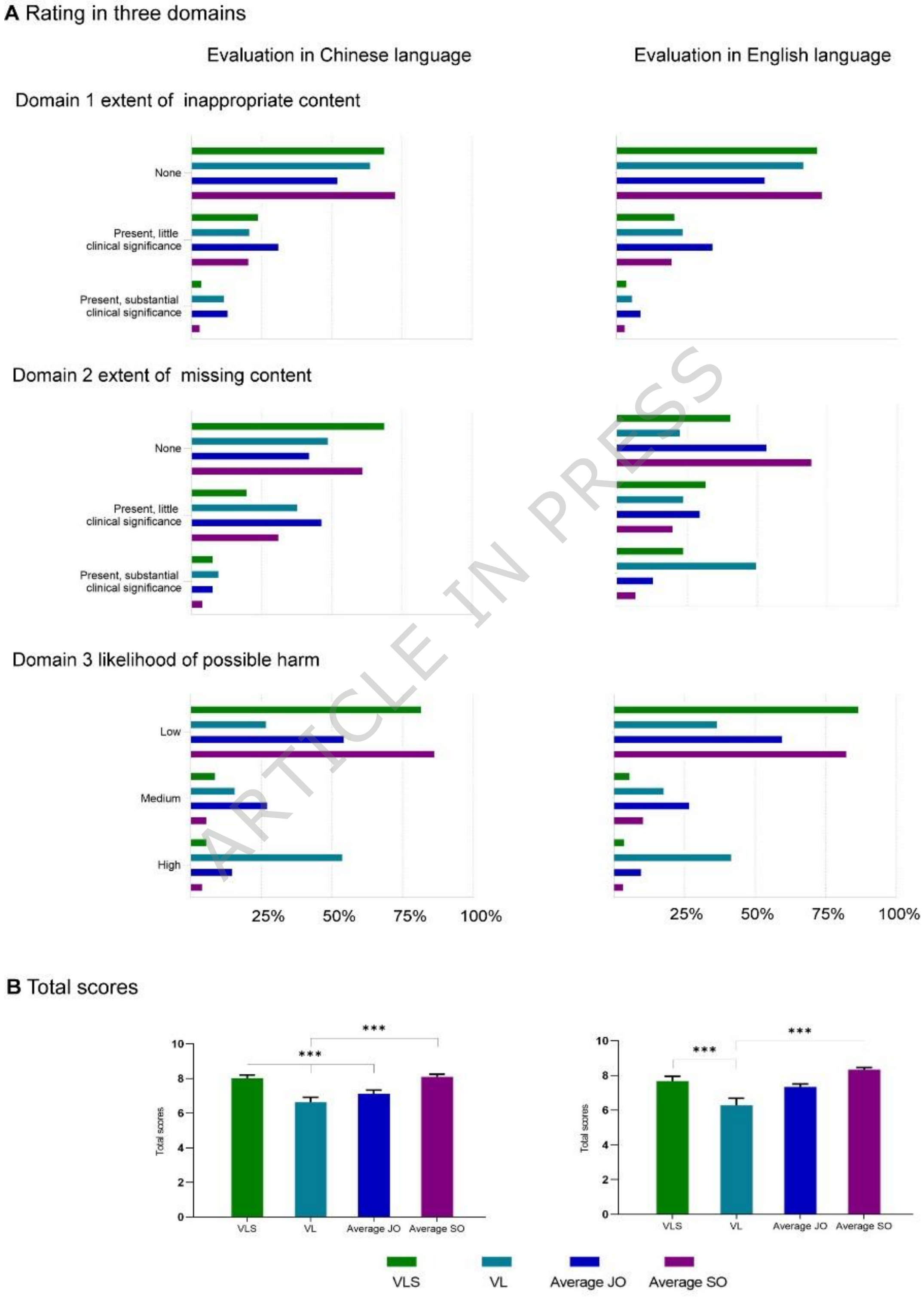


Figure 6. Comparison of VLS (Vision-Language Segmentation) models, VL, 3 junior ophthalmologists

(JO), and 3 senior ophthalmologists (SO) in both English and Chinese. (A) Evaluators rated management recommendations across three criteria: appropriateness, completeness, and potential harm, using 100 cases. (B) Total scores of management recommendations by VLS, VL, JO, and SO based on 100 cases. Box plot (n = 100), showing median, quartiles, and data range (whiskers). Comparisons were conducted using two-sided Friedman tests, with post-hoc pairwise comparisons using two-sided Wilcoxon signed-rank tests. P-values for multiple comparisons were adjusted using the Bonferroni method. *p < 0.05, **p < 0.01, and ***p < 0.001.

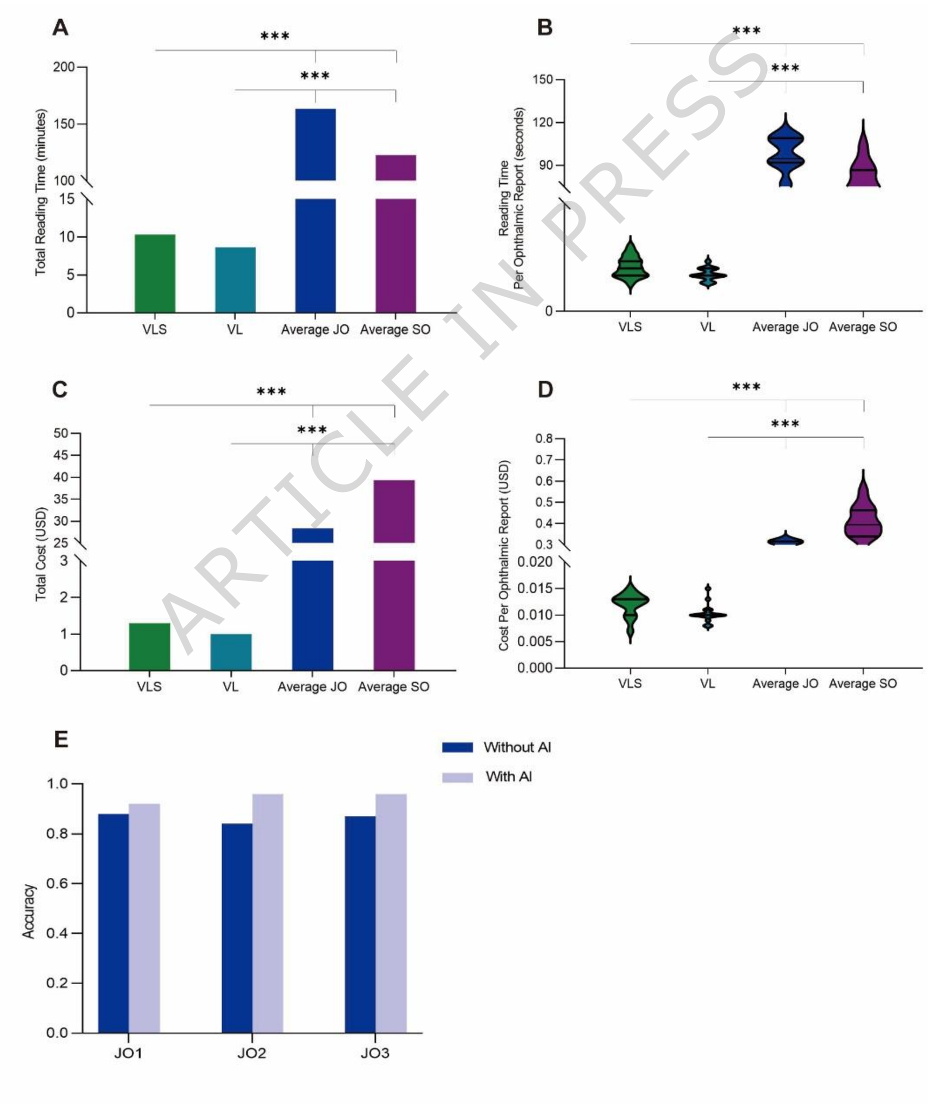


Figure 7. (A) Bar graph displaying total reading time in seconds for VLS (Vision-Language Segmentation Models), VL, 3 junior ophthalmologists (JO), and 3 senior ophthalmologists (SO), (B) Violin plot illustrating reading time per ophthalmic report in seconds, (C) Bar graph presenting total cost in U.S. dollars for VLS, VL, and human readers, (D) Violin plot showing cost per ophthalmic report in U.S. dollars. Dashed lines represent medians and dotted lines indicate quartiles, and (E) diagnostic accuracy of three junior ophthalmologists with and without AI assistance across 200 ophthalmic ultrasound cases. ***p < 0.001.

Tables

Table 1. Demographic data of the study and distribution of the findings.

Items	SAHZU Dataset			FAHWM Dataset	FAHZC Dataset	Total
	Training set	Validation set	Test set	External test set 1	External test set 2	
Patients, n	5497	1915	1919	269	70	9670
Images, n	37917	12639	12640	742	160	64098
Reports, n	12649	4197	4170	269	70	21355
Age, mean (SD)	49.6 (17.2)	49.5 (17.4)	49.7 (17.2)	50.8 (18.2)	57.4 (16.1)	51.4 (17.2)
Gender, n (%)						
Male	2615 (47.6)	906 (47.3)	904 (47.1)	108 (40.1)	32 (45.7)	4565 (47.2)
Female	2882 (52.4)	1009 (52.7)	1015 (52.9)	161 (59.9)	38 (54.3)	5105 (52.8)
Eye, n (%)						
OS	2905 (23)	996 (23.7)	978 (23.5)	NA	17 (24.2)	4896 (22.9)
OD	3160 (25)	1068 (25.5)	1068 (25.6)	NA	16 (22.9)	5312 (24.9)
OS&OD	6584 (52)	2133 (50.8)	2124 (50.9)	269 (100)	37 (52.9)	11147 (52.2)
Report length, mean (SD)	99.6 (33.8)	99.7 (34)	99.6(33.6)	28.6 (9.3)	132.8(41.1)	92.1 (30.4)
Diagnosis, n (%)						
Cataract	1350 (10.7)	445 (10.6)	454 (10.9)	32 (11.9)	10 (14.2)	2291 (10.7)
VH	3016 (23.8)	1123 (26.8)	955 (22.9)	29 (10.8)	10 (14.3)	5133 (24.0)
HM	2336 (18.5)	763 (18.2)	794 (19.1)	23 (8.6)	10 (14.3)	3926 (18.5)
UM	16 (0.1)	4 (0.1)	6 (0.1)	15 (5.6)	10 (14.3)	51 (0.2)
RE	1711 (13.6)	491 (11.7)	495 (11.9)	105 (39.0)	10 (14.3)	2812 (13.2)
RB	15 (0.1)	6 (0.1)	5 (0.1)	30 (11.1)	10 (14.3)	66 (0.3)
RD	4205 (33.2)	1365 (32.5)	1461 (35)	35 (13.0)	10 (14.3)	7076 (33.1)

VH = vitreous hemorrhage, HM = high myopia, UM = uveal melanoma, RE = refractive error, RB = retinoblastoma, RD = retinal detachment. The Second Affiliated Hospital of Zhejiang University (SAHZU), the First Affiliated Hospital of Zhejiang Chinese Medical University (FAHZC), and the First Affiliated Hospital of Wannan Medical College (FAHWM).