IJHSR
ijhighschoolresearch.org
RESEARCH ARTICLE

A Comparative Study of Large Vision-Language Models for Retinal Image-Based Alzheimer's Disease Screening

Vincent Z Wang, Yiyi Sun

Brophy College Preparatory, 4701 N. Central Ave. Phoenix, AZ, 85012, USA; vzwang33@gmail.com
Mentor: Oana M Dumitrascu MD, MSc

ABSTRACT: Retinal imaging has emerged as a promising non-invasive approach for early Alzheimer's disease (AD) screening, offering a window into neurodegeneration through accessible ocular biomarkers. Concurrently, whereas large vision-language models (VLMs) have shown remarkable performance across general visual reasoning tasks, their diagnostic potential for medical imaging remains underexplored. In this study, we systematically appraised five VLMs—LLaMA, LLaVA, LLaVA-Med, Qwen, and RetinalGPT—for detecting AD from retinal color fundus photographs (CFPs). Using a curated subset of retinal fundus images from the UK Biobank (114 CFPs from 107 AD subjects and 115 CFPs from 105 cognitively normal controls), we assessed diagnostic reasoning, accuracy, uncertainty ratio, and interpretability. RetinalGPT, fine-tuned on ophthalmic datasets, outperformed general-purpose models with higher classification accuracy (59.9%) and F1 score (0.45) while maintaining greater medical specificity and confidence. In contrast, general VLMs displayed inconsistent reasoning and excessive uncertainty (up to 5.34%). Our findings demonstrate the importance of domain-specific fine-tuning of VLMs for routine clinical applications and propose a hybrid inference framework leveraging uncertainty-based model selection to balance scalability and screening accuracy.

KEYWORDS: Medical and Health Sciences, Computer Science and Software Engineering, Artificial Intelligence, Retinal Imaging, and Alzheimer's Disease.

Introduction

Alzheimer's disease (AD) is the most prevalent form of dementia, affecting more than 55 million people globally.[1] Early identification of individuals at risk of AD has become a top scientific and public health priority, as timely interventions during preclinical or mild cognitive impairment stages can mitigate disease progression and generate substantial long-term healthcare savings.[2] Biomarkers of AD, including amyloid (A), tau (T), neurodegeneration (N), and inflammation (I), provide important diagnostic insights,[3] yet their measurement typically requires invasive procedures such as lumbar puncture or costly neuroimaging modalities like magnetic resonance imaging (MRI) and positron emission tomography (PET). These limitations underscore the pressing need for noninvasive, scalable, and affordable biomarkers that can detect AD-related changes earlier and more broadly across diverse populations.

The retina, a direct and accessible extension of the central nervous system (CNS), offers a promising window into neurodegenerative disorders. *In vivo* and post-mortem studies have demonstrated that retinal neurodegeneration and microvascular alterations closely mirror cerebral Alzheimer's pathology.[4,5] Retinal color fundus photography (CFP) enables the safe, rapid, and cost-effective visualization of these microstructural and vascular features, making it highly suitable for screening in primary care, ophthalmology, and community settings. Such accessibility positions retinal imaging as an emerging surrogate biomarker for AD, providing a method to investigate vascular and neurodegenerative changes associated with early AD.[6,7]

Concurrently, advances in large vision–language models (VLMs) have transformed artificial intelligence (AI) research by combining visual perception with natural language reasoning.[8] Originating from transformer architectures, modern VLMs such as CLIP, BLIP-2, Flamingo, LLaVA, Qwen-VL, and GPT-4V merge image interpretation and linguistic understanding, enabling models to describe, reason, and converse about visual content. These systems have achieved state-of-the-art performance in tasks including autonomous robotics, education, biomedical research, and scientific data interpretation, demonstrating the remarkable power of multimodal alignment between vision and language. Building upon this paradigm, multimodal foundation models are now reshaping computational medicine, supporting tasks such as radiology report generation, pathology slide interpretation, dermatologic triage, and ophthalmic screening.[9–11]

In ophthalmology, the field has recently witnessed transformative work in domain-specific vision–language modeling. RETFound, a model trained on more than 1.6 million retinal images, demonstrated that foundation models pretrained on large ophthalmic datasets can generalize to diverse downstream tasks, including diabetic retinopathy and glaucoma detection.[12] Extending this approach, EyeFM, a multimodal eyecare foundation model integrating 14.5 million ocular images across five modalities with paired clinical texts, achieved substantial improvements in diagnostic accuracy and clinical report standardization across multinational cohorts and a double-masked randomized controlled trial.[13] These studies collectively illustrate how domain-tuned VLMs can augment clinicians' diagnostic accuracy and reasoning transparency, particularly when jointly trained on multimodal imaging and structured textual data.

DOI: 10.36838/IJHSR812.62

Nevertheless, despite these advances, the translation of general-purpose VLMs into clinical diagnostics remains challenging. Models trained predominantly on natural images and conversational text often lack medical specificity, consistent factual grounding, and interpretable reasoning. Current evidence suggests that clinical deployment demands specialized instruction tuning and multimodal alignment with domain-specific ontologies to ensure safe, explainable, and equitable performance.[13–15] Addressing this gap, our study explores how general and domain-specialized VLMs differ in their ability to reason over retinal CFPs, offering new insights into the frontier between general artificial intelligence and clinical expertise.

This study addresses translational gaps by conducting a comparative evaluation of five representative VLMs, spanning general-purpose architectures and domain-specialized clinical assistants, to assess their ability to perform diagnostic reasoning on CFPs. Specifically, the models include LLaMA 2.2–11B–Vision, LLaVA, LLaVA-Med, Qwen-VL, general-purpose models, and RetinalGPT, a domain-specific, instruction-tuned ophthalmic model. Through this framework, we aim to quantify diagnostic accuracy and reasoning reliability across general and domain-tuned VLMs.

Our central hypothesis is that *specialized, instruction-tuned VLMs* trained on ophthalmic datasets will demonstrate comparatively stronger diagnostic accuracy, domain coherence, and interpretability compared to general-purpose VLMs. We posit that aligning multimodal representations with retinal vascular and morphological biomarkers enables finer discrimination of disease-related features and more clinically intelligible reasoning. Retinal imaging thus serves as an ideal testbed for evaluating whether modern multimodal AI systems can transcend pattern recognition and perform clinically meaningful, explainable inference in AD screening.

Methods

Dataset:

The dataset comprised 229 retinal CFPs from the UK Biobank[16] (http://www.ukbiobank.ac.uk/about-biobank-uk), including subjects with AD (114 CFPs from 107 patients) and without AD (115 CFPs from 105 subjects). The AD label was based on International Classification of Diseases (ICD) codes from hospital admission and death records, indicating a definitive clinical diagnosis of dementia caused by AD (Data-field 42021). The non-AD label was based on the absence of neurodegenerative conditions and other dementias.[7]

After image quality control, the dataset primarily included one retinal CFP per subject; however, CFPs from both eyes were retained for a subset of participants, resulting in paired images for 10 non-AD subjects and 7 AD subjects. Each CFP was treated as an independent image-level input to the VLMs, consistent with common practice in ophthalmic imaging studies where retinal labels are typically assigned at the eye level rather than the subject level.[6,7,17]

For each input image, we first detected the retina using the Hough circle transform and then cropped the mask region to minimize the effect of the black background. Afterward, the images were resized to 224 × 224 and normalized to (–1, 1). Special efforts were made to ensure that none of the CFPs in this subset were used in training RetinalGPT. We evaluated five VLMs: LLaMA, LLaVA, LLaVA-Med, Qwen, and RetinalGPT. Each model received identical prompts describing CFPs and was asked to classify each as AD positive, negative, or uncertain. Performance was evaluated using accuracy, F1 score, uncertainty ratio (UR), and reasoning score (RS). An expert neuro-ophthalmologist (OD) qualitatively reviewed a subset of outputs to assess accuracy.

Model Selection:

We evaluated five VLMs representing a spectrum from general-purpose multimodal reasoning to ophthalmology-specific fine-tuning. Each model was selected to capture a distinct axis of capability: from text-only baselines extended with vision encoders to specialized models trained on retinal data. This diversity allows for an assessment of how domain alignment and multimodal instruction tuning affect medical reasoning performance.

1. Llama 3.2 11B Vision (Large Language Model Meta AI):[18]

Llama 3.2 11B Vision serves as a text-centric baseline augmented with a vision adapter layer. Derived from Meta AI's LLaMA ^.2 family, the 11B parameter version integrates frozen CLIP-L/14 visual embeddings into the transformer's token stream through a low-rank adapter (LoRA) projection. While LLaMA ^.2 itself is designed for instruction-following and text generation, the visual encoder enables rudimentary image captioning and multimodal grounding. However, this architecture remains largely non-specialized for medical imagery, lacking domain-specific feature calibration. Its inclusion in this study highlights the general language-vision reasoning ability of a large but generic foundation model when confronted with CFPs.

2. LLaVA (Large Language and Vision Assistant):[19]

LLaVA represents a general-purpose open-source multimodal model that tightly couples a visual encoder (CLIP-ViT-L/14) with a large language model (Vicuna-13B or LLaMA backbone). Through *visual instruction tuning*, a process aligning image–text representations with conversational supervision, LLaVA achieves strong zero-shot performance across reasoning, captioning, and visual question answering tasks. Its multimodal training corpus spans over 600,000 human-annotated instruction–response pairs from diverse image domains, providing robust grounding for scene understanding and reasoning tasks. LLaVA's architecture allows it to handle fine-grained image–text alignment, but since its data are primarily non-medical, it is expected to generalize broadly but not deeply in medical interpretation. In this study, it serves as a high-performing general VLM against which domain-tuned models can be compared.

3. LLaVA-Med: [20]

LLaVA-Med extends LLaVA's general architecture into the biomedical and clinical imaging domain through multistage

instruction tuning on domain-specific corpora such as PMC-15M, PubMedQA, and MedPix. Following the approach described by Li *et al.*,[20] LLaVA-Med integrates approximately 60K medical image–text pairs covering modalities including radiology, pathology, endoscopy, and ophthalmology. The model learns to follow diagnostic-style instructions, describe anatomical structures, and justify findings in clinical language. Its training follows a two-stage strategy: (1) *alignment fine-tuning* to map visual features from biomedical images to the pretrained language space, and (2) *instruction tuning* to learn task-specific reasoning behaviors. LLaVA-Med thus provides a strong medical-domain baseline with broader biomedical understanding but limited specialization for retinal microvascular patterns or AD phenotyping.

4. Qwen-VL: [21]

Developed by Alibaba Cloud, Qwen-VL is a compact and computationally efficient multimodal model designed for scalable deployment. It combines a Vision Transformer (ViT-Large/14) with the Qwen-7B or 14B language backbone through a cross-attention fusion layer. Despite its lighter architecture, Qwen-VL demonstrates competitive performance on multiple benchmarks such as ScienceQA, MME, and OCR-VQA. Its key strength lies in its dense *vision–language fusion* that supports fine-grained perception of localized features, which is beneficial for analyzing vascular morphology in fundus images. While not tuned on biomedical data, its strong visual attention mechanisms make it a promising general model for image-anchored reasoning and an efficient candidate for downstream clinical adaptation.

5. RetinalGPT: [22]

RetinalGPT represents a domain-specific adaptation built for ophthalmic imaging and AD-related vascular biomarker interpretation. Building upon an LLaVA-Med base, RetinalGPT underwent *visual instruction tuning* on 38K CFP images from multiple ophthalmology datasets, including Messidor-1,[23] APTOS,[24] EyeQ,[25] IDRiD,[26] MICCAI MACC,[27] OIA-ODIR,[28] RFMiD,[29] and UK Biobank[16] (no overlap with our evaluation dataset). The tuning pipeline incorporated structured instruction–response pairs emphasizing quantitative vascular features, such as fractal dimension, vessel tortuosity, and branching angle. This specialized alignment allowed the model to learn clinically preferred reasoning patterns, e.g., "increased retinal vessel tortuosity and reduced fractal dimension correlate with neurodegenerative progression."

RetinalGPT also employs multi-turn dialogue supervision, enabling it to generate coherent, stepwise diagnostic rationales rather than single-sentence outputs. This makes it uniquely suited for *interpretable medical reasoning* in retinal image analysis. In this study, it functions as a specialized and fine-tuned model, providing insight into how targeted instruction tuning improves domain fidelity and interpretability relative to general-purpose VLMs.

Comparative Summary and Conceptual Contrast:

Collectively, the five selected VLMs represent a continuum from general foundation systems to domain-specialized clinical assistants, enabling a nuanced evaluation of multimodal reasoning transfer.

At one end of the spectrum, *LLaMA 2.2–11B–Vision* and *LLaVA* embody the general-purpose paradigms that rely on large-scale, non-clinical instruction tuning to achieve versatility and compositional reasoning. Their strength lies in open-ended visual understanding, but their clinical inferences tend to be descriptive rather than diagnostic, reflecting a lack of exposure to medically grounded data distributions.

In contrast, *LLaVA-Med* exemplifies the biomedical intermediary, a bridge model that leverages large biomedical corpora (e.g., PMC-15M) to achieve robust cross-modal alignment within the clinical lexicon. It demonstrates strong transfer learning potential for generic medical image interpretation but remains limited in fine-grained morphological sensitivity, such as the retinal microvascular abnormalities associated with AD retinal degeneration.

Qwen-VL occupies a unique position in this taxonomy: though trained on general data, its dense vision–text fusion architecture and efficient cross-attention design enable high spatial precision, making it competitive with larger models despite lower parameter counts. Its performance provides insight into how architectural efficiency can partially compensate for the absence of domain adaptation.

Finally, *RetinalGPT* represents the domain-specialized end of the spectrum. By fusing ophthalmic imaging datasets with structured vascular descriptors and disease-relevant instructions, *RetinalGPT* learns to emulate expert clinical reasoning patterns. It does not merely classify but can articulate *why* a retinal feature implies a diagnostic outcome, generating transparent justifications aligned with human clinical logic. This makes it an effective benchmark for *interpretable medical AI*, a capability essential for real-world deployment in healthcare settings.

Together, these five models enable a systematic assessment of how instruction-tuning depth, dataset specificity, and architectural design influence diagnostic performance, uncertainty behavior, and reasoning coherence in multimodal medical contexts. By positioning RetinalGPT against its general and semi-specialized counterparts, this study isolates the contributions of domain alignment to both quantitative accuracy and qualitative interpretability, shedding light on the evolving boundary between general intelligence and clinical precision. Table 1 shows an overview of five evaluated VLM models.

All models were deployed in a secure offline environment to ensure data confidentiality. No external API access was used.

DOI: 10.36838/IJHSR812.62

Table 1: Overview of Evaluated Vision–Language Models. The table summarizes model architecture, training data, instruction-tuning strategies, and specialization level, spanning general foundation systems to the retinal domain-specific RetinalGPT. This comparison underscores how increasing domain alignment, from generic to ophthalmology-tuned models, progressively improves medical interpretability and diagnostic potential.

Model Name	Training Data Source	Instruction Tuning Strategy	Parameter Scale	Domain Scope	Specialization Level
LLaMA 2.2–11B–Vision	Generic web-scale multimodal data; LAION-2B and synthetic caption datasets	Minimal: visual adapter alignment only	11 B	General	Text-centric baseline
LLaVA	600 K human-annotated multimodal instruction–response pairs	Visual Instruction Tuning on general images	13 B	Multidomain general	Multimodal generalist
LLaVA-Med	PMC-15M, MedPix, PubMedQA	Two-stage tuning: image–text alignment and clinical instruction tuning	13 B	Biomedical	Medical domain generalist
Qwen-VL	General visual corpus (Visual Genome, COCO, MME)	Cross-attention multimodal alignment; weak supervision	7–14 B	Multidomain	Lightweight generalist
RetinalGPT	Messidor-1, APTOS, EyeQ, IDRiD, MICCAI MACC, OIA-ODIR, RFMiD, UK Biobank	Multistage visual instruction tuning on retinal datasets; clinical reasoning supervision	≈ 13 B (+LoRA adapters)	Ophthalmology	Domain-specialized (retinal)

Experimental Procedures:

Retinal CFPs were standardized to 512 x 512 resolution and normalized to (-1, 1) following the protocol described in a prior study.[7] The experiment pipeline was automated with deterministic seeding for reproducibility. In the experiments, each VLM model received the same multimodal input prompts containing:

1. A two-dimensional retinal image (CFP).
2. A concise natural-language description (patient age, condition context).
3. A diagnostic query, e.g.:

"Based on this retinal image, is there evidence of Alzheimer-related microvascular abnormalities?"

Each model received identical multimodal prompts describing retinal images and was instructed to classify each case as:

- Positive (indicative of AD),
- Negative (normal), or
- Uncertain (requires further evaluation).

The "Uncertain" classification was operationalized using a rule-based analysis of the models' textual outputs. An explicit classification indicating either positive for Alzheimer's disease (AD) or negative for AD (cognitively normal) was required for a definitive label. Model responses that did not clearly indicate either a positive or negative AD classification, such as those expressing insufficient information, diagnostic caution, or contextual discussion without a conclusion, were classified as "Uncertain."

To ensure interpretability, models were prompted to generate both structured classification outputs and free-text justifications, allowing qualitative assessment of diagnostic reasoning.

Responses were parsed for structured classification outputs and free-text justifications. The following metrics were computed:

- Accuracy (ACC): proportion of correctly classified images divided by the total number of images.
- F1 Score (F1): harmonic mean of precision and recall, highlighting the balance between sensitivity and specificity.
- Uncertainty Ratio (UR): proportion of cases labeled as *Uncertain*, reflecting the model's epistemic uncertainty and diagnostic confidence.
- Reasoning Score (RS): expert-rated scale (1–5) evaluating medical coherence and explanation quality. It was automatically evaluated using Mistral-7B.[30]

For performance evaluation, cases classified as "Uncertain" were excluded from accuracy and F1 score calculations, which were computed only over images receiving definitive positive or negative AD classifications. This approach avoids penalizing models for appropriately abstaining in ambiguous scenarios while enabling fair comparison of diagnostic performance.

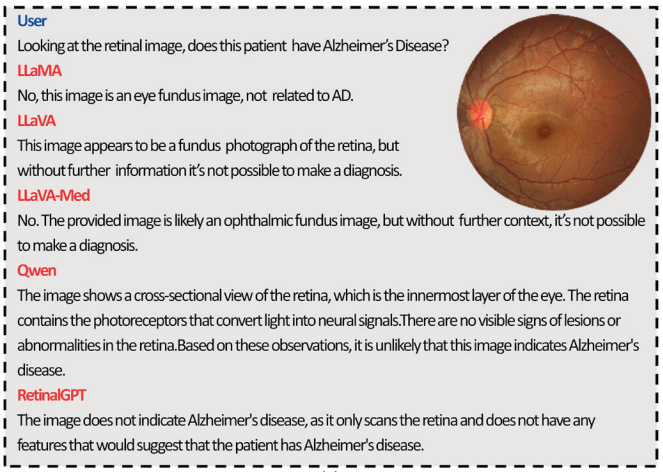
Reasoning Score (RS) quantified the clinical relevance and diagnostic usefulness of model-generated explanations on a 1–5 scale. To enable scalable evaluation, RS was automatically estimated using the Mistral-7B language model, which was prompted to act as a biomedical expert and score each explanation according to a predefined rubric (1 = clinically irrelevant or uninformative; 3 = partially relevant but weak or uncertain; 5 = highly relevant and informative for disease assessment).

To contextualize and calibrate this automated evaluation, we also qualitatively reviewed representative model outputs and corresponding RS values to verify that higher scores reflected clinically coherent and diagnostically meaningful reasoning. This expert-in-the-loop assessment confirmed that the automated scoring captured relative differences in explanation quality across models. Accordingly, RS is intended as a comparative proxy for reasoning quality rather than a substitute for formal expert rating or clinical validation.

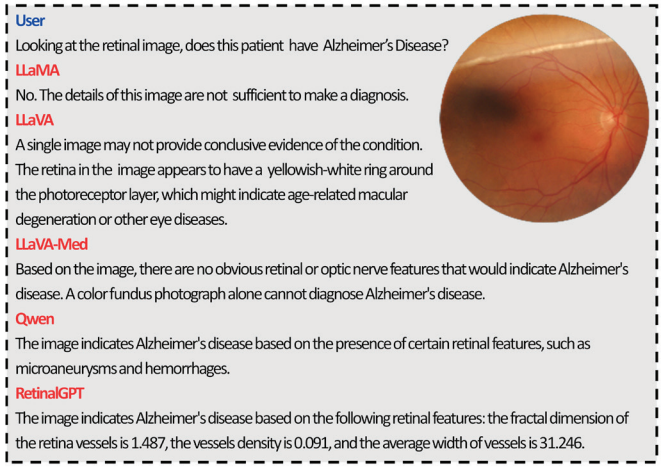
Results and Discussion

Representative responses generated by the five evaluated VLMs when asked to assess whether a retinal CFP indicates AD pathology are illustrated in Figure 1 (the top row (a) shows an image of a normal control (NC) subject, and the bottom row (b) shows an image of an AD patient). The example highlights distinct reasoning behaviors and diagnostic limitations across models. The general-purpose LLaMA and LLaVA models recognized the image as a retinal CFP but failed to provide medically contextualized interpretations. LLaVA-Med, although trained on biomedical data, correctly identified the ophthalmic nature of the image, yet expressed diagnostic uncertainty due to insufficient contextual information. Qwen-VL demonstrated more structured reasoning by describing retinal anatomy and explicitly concluding the absence of AD-related lesions, though its explanation remained general. In contrast, the domain-specific RetinalGPT produced the most clinically intelligible output, acknowledging the image's modality, its diagnostic scope, and explicitly noting the absence of AD biomarkers such as vascular irregularities. Collectively, these outputs demonstrate the progressive improvement in domain awareness and reasoning specificity from general-purpose to ophthalmology-tuned models, underscoring the value of tar-

geted instruction tuning for clinically reliable multimodal inference.



(a)



(b)

Figure 1: Representative qualitative responses from five vision–language models (LLaMA, LLaVA, LLaVA-Med, Qwen-VL, and RetinalGPT) to a retinal CFP classification prompt. Each model was asked to assess whether a given CFP was positive for AD, negative for AD, or uncertain, and to provide a brief explanatory rationale. Panel (a) shows responses for an image from a cognitively normal control subject, and panel (b) shows responses for an image from an AD subject. These examples highlight that domain-specific tuning markedly enhances medical reasoning quality and the clinical relevance of generated explanations.

Figure 2: Performance comparison of five vision–language models on AD retinal image classification and reasoning tasks using 229 retinal color fundus photographs (114 images from AD patients and 115 from cognitively normal controls). Models were prompted to classify each image as positive for AD, negative for AD, or uncertain. Reported metrics include uncertainty ratio (UR), accuracy (ACC), F1 score, and reasoning score (RS). RS reflects rubric-guided automated evaluation of explanation quality with expert oversight. Results illustrate relative differences in model behavior under this experimental framework. RetinalGPT achieved the highest accuracy (59.9%), lowest uncertainty, and strongest reasoning coherence, confirming that ophthalmic fine-tuning substantially improves diagnostic reliability over general-purpose models.

Model	Uncertainty Ratio (UR)	Accuracy (ACC)	F1	Reasoning Score (RS)
LLaMA	5.34%	53.90%	0.21	1.76±0.97
LLaVA	3.49%	48.42%	0.17	2.51±1.11
LlaVA-Med	2.18%	51.79%	0.56	3.28±1.11
Qwen	1.75%	47.56%	0.02	1.60±0.92
RetinalGPT	0.8%	59.91%	0.45	3.54±1.57

The quantitative evaluation of the five VLMs across diagnostic accuracy (ACC), uncertainty ratio (UR), F1 score, and expert-rated reasoning quality (RS) is summarized in Table 2. A consistent performance gradient emerged between general-purpose and domain-specialized models.

Among general-purpose models, LLaMA achieved the strongest overall baseline, with an accuracy of 53.9%, an F1 score of 0.21, and an RS of 1.76 ± 0.97, despite a moderate UR of 5.34%. This performance suggests that even non-domain-tuned VLM frameworks can internalize rudimentary diagnostic cues when prompted effectively. LLaVA, another general-purpose model with visual alignment, exhibited slightly higher uncertainty (3.49%) and lower accuracy (48.42%), but generated more descriptive, albeit less clinically precise, explanations (RS = 2.51 ± 1.11).

Introducing medical fine-tuning led to a noticeable improvement. LLaVA-Med demonstrated both higher diagnostic precision (ACC = 51.79%) and markedly enhanced interpretability (RS = 3.28 ± 1.11), while maintaining a reduced UR (2.18%). Its F1 score (0.56) indicates that biomedical adaptation improves both confidence and semantic alignment with disease-specific retinal features, though subtle domain mismatches remain visible.

The Qwen-VL model, optimized for general-purpose multimodal reasoning but without explicit clinical exposure, displayed the lowest uncertainty among all base models (1.75%) but also the lowest F1 score (0.02) and modest reasoning coherence (RS = 1.60 ± 0.92). This suggests that while Qwen-VL achieves stable confidence calibration, it lacks domain-relevant contextual grounding for medical inference.

By contrast, the RetinalGPT model—explicitly tuned on ophthalmic and neurodegenerative imaging data—achieved the best overall performance. It produced the highest accuracy (59.91%), lowest uncertainty (0.8%), and a strong reasoning score of 3.54 ± 1.57, reflecting both robust diagnostic reliability and consistent medical interpretability. RetinalGPT's F1 score (0.45) further supports its balanced precision–recall behavior, distinguishing it as the most reliable model across both diagnostic and explanatory dimensions.

To assess within-subject consistency, we examined VLM predictions across paired CFPs from subjects for whom images from both eyes were available (10 NC subjects and 7 AD subjects). Predictions across paired CFPs showed varying degrees of concordance across models, with domain-specialized models exhibiting greater sensitivity but also increased variability relative to general-purpose models.

Qualitative analysis further reinforced these trends. RetinalGPT generated concise, pathology-aware explanations referencing features such as optic disc pallor, vessel attenuation, and retinal texture abnormalities, patterns consistent with AD–related retinal neurodegeneration. In comparison, general-purpose models like LLaVA and LLaVA-Med produced visually coherent but diagnostically ambiguous descriptions, while Qwen-VL often misattributed disease likelihood to non-medical color or texture variations.

Furthermore, disparities in reasoning scores across models indicate that general-purpose VLMs struggle when confront-

DOI: 10.36838/IJHSR812.62

ed with unfamiliar medical terminology or subtle retinal biomarkers, even when producing fluent textual descriptions. In contrast, domain-specialized models achieved higher reasoning scores by generating explanations that were more clinically grounded and internally consistent. This gap in reasoning quality highlights the importance of domain-specific instruction tuning for reliable medical interpretation and motivates architectures that selectively leverage specialized models when higher-quality reasoning is required.

Across models, uncertainty inversely correlated with accuracy ($r = -0.76$), indicating that uncertainty can serve as a surrogate indicator of diagnostic confidence. This suggests potential for hierarchical inference pipelines—deploying lightweight general models for initial triage and invoking specialized models (e.g., RetinalGPT) when low-confidence predictions arise, thereby optimizing computational efficiency while preserving diagnostic fidelity.

Taken together, these findings indicate that general-purpose VLMs, though proficient at image-language alignment, lack the precision needed for clinical interpretation. Their tendency to produce verbose but vague explanations suggests a deficiency in medically grounded feature understanding. RetinalGPT's comparatively stronger performance highlights the importance of domain-specific instruction tuning and task-aware alignment for medical reasoning. These results parallel prior findings in biomedical natural language processing (NLP), where adaptation to specialized corpora substantially improves factual accuracy.[31]

Importantly, within-subject variability across CFPs does not necessarily indicate poor model performance. Retinal images are commonly labeled and analyzed at the eye level, particularly in large-scale ophthalmic studies such as those focused on diabetic retinopathy, where disease manifestations may differ between eyes. In this study, labels were assigned at the subject level, which may introduce apparent inconsistencies when eye-specific retinal features differ. As such, observed discordance across paired CFPs should be interpreted as an exploratory finding reflecting biological asymmetry, imaging variability, and labeling granularity, rather than definitive evidence of model error.

Although the absolute classification accuracies of the evaluated VLMs are modest, prior studies using conventional convolutional neural networks and deep learning pipelines on retinal fundus photographs have reported substantially higher performance for AD detection under supervised learning settings.[7,17] This suggests that the task itself is tractable with task-specific architectures, and that the observed performance limitations primarily reflect current constraints of general-purpose and instruction-tuned VLMs rather than intrinsic dataset difficulty. Importantly, the goal of this study was not to establish a state-of-the-art classifier or directly compare VLMs with optimized traditional machine learning methods, but to assess whether domain-specific instruction tuning improves diagnostic reasoning, uncertainty behavior, and interpretability relative to general-purpose multimodal models under identical prompting conditions. The significance of VLM-based approaches, therefore, lies not in immediate competitive accuracy, but in their ability to integrate image interpretation, uncertainty expression, and natural-language explanation within a unified framework—capabilities that remain difficult to achieve with conventional classifiers.

Limitations and Future Work

A relatively limited CFP dataset was used for these models' performance determination. In particular, the relatively small sample size may affect the stability of uncertainty estimates and cross-model comparisons; therefore, differences in uncertainty ratios or overall performance should be interpreted with caution in the context of potential clinical applicability. While RetinalGPT achieved the highest performance among the evaluated models, its accuracy remains below the threshold required for autonomous diagnostic use.

Future work will include fine-tuning on larger, multi-center AD-derived datasets of CFPs, integrating CNN backbones for hybrid inference, and developing explainable AI visualizations to further enhance interpretability. It will also be important to investigate eye-specific labeling strategies and larger cohorts with paired CFPs to better disentangle biological asymmetry from model uncertainty and to refine subject-level inference. In addition, extending the comparative framework to foundation models like GPT-4V and Med-Flamingo will enable broader benchmarking. Finally, model interpretability should be evaluated through clinical user studies to assess trustworthiness and adoption feasibility.

Conclusion

This comparative study demonstrates that while VLMs hold immense potential for medical imaging analysis, domain specialization remains essential for diagnostic reliability and interpretability. RetinalGPT's enhanced performance highlights the benefits of biomedical fine-tuning, whereas general-purpose models exhibit high uncertainty and limited medical explainability. Importantly, this study should be viewed as an exploratory investigation into the behavior of vision–language models on retinal images, rather than a demonstration of a deployable diagnostic system. Future work will include direct comparisons with traditional deep learning baselines under matched data splits, as well as the development of hybrid frameworks that leverage uncertainty-aware routing to better characterize trade-offs between classification accuracy, multimodal reasoning capabilities, and scalable, clinically aligned AI systems. This work contributes foundational insights toward deploying multimodal foundation models in sensitive clinical domains such as AD screening.

Acknowledgments

This research has been conducted using the UK Biobank Resource under Application Number 91706.

References

1. Brookmeyer, R.; Johnson, E.; Ziegler-Graham, K.; Arrighi, H. M. Forecasting the Global Burden of Alzheimer's Disease. *Alz-*

heimers Dement **2007**, *3* (3), 186–191. https://doi.org/10.1016/j.jalz.2007.04.381.
2. Rasmussen, J.; Langerman, H. Alzheimer's Disease - Why We Need Early Diagnosis. *Degener. Neurol. Neuromuscul. Dis*. **2019**, *9*, 123–130. https://doi.org/10.2147/DNND.S228939.
3. Jack, C. R., Jr.; Bennett, D. A.; Blennow, K.; Carrillo, M. C.; Feldman, H. H.; Frisoni, G. B.; Hampel, H.; Jagust, W. J.; Johnson, K. A.; Knopman, D. S.; Petersen, R. C.; Scheltens, P.; Sperling, R. A.; Dubois, B. A/T/N: An Unbiased Descriptive Classification Scheme for Alzheimer Disease Biomarkers. *Neurology* **2016**, *87* (5), 539–547. https://doi.org/10.1212/WNL.0000000000002923.
4. Gaire, B. P.; Koronyo, Y.; Fuchs, D.-T.; Shi, H.; Rentsendorj, A.; Danziger, R.; Vit, J.-P.; Mirzaei, N.; Doustar, J.; Sheyn, J.; Hampel, H.; Vergallo, A.; Davis, M. R.; Jallow, O.; Baldacci, F.; Verdooner, S. R.; Barron, E.; Mirzaei, M.; Gupta, V. K.; Graham, S. L.; Tayebi, M.; Carare, R. O.; Sadun, A. A.; Miller, C. A.; Dumitrascu, O. M.; Lahiri, S.; Gao, L.; Black, K. L.; Koronyo-Hamaoui, M. Alzheimer's Disease Pathophysiology in the Retina. *Prog. Retin. Eye Res*. **2024**, *101*, 101273. https://doi.org/10.1016/j.preteyeres.2024.101273.
5. Mirzaei, N.; Shi, H.; Oviatt, M.; Doustar, J.; Rentsendorj, A.; Fuchs, D.-T.; Sheyn, J.; Black, K. L.; Koronyo, Y.; Koronyo-Hamaoui, M. Alzheimer's Retinopathy: Seeing Disease in the Eyes. *Front. Neurosci*. **2020**, *14*.
6. Cheung, C. Y.; Ran, A. R.; Wang, S.; Chan, V. T. T.; Sham, K.; Hilal, S.; Venketasubramanian, N.; Cheng, C.-Y.; Sabanayagam, C.; Tham, Y. C.; Schmetterer, L.; McKay, G. J.; Williams, M. A.; Wong, A.; Au, L. W. C.; Lu, Z.; Yam, J. C.; Tham, C. C.; Chen, J. J.; Dumitrascu, O. M.; Heng, P.-A.; Kwok, T. C. Y.; Mok, V. C. T.; Milea, D.; Chen, C. L.-H.; Wong, T. Y. A Deep Learning Model for Detection of Alzheimer's Disease Based on Retinal Photographs: A Retrospective, Multicentre Case-Control Study. *Lancet Digit. Health* **2022**, *4* (11), e806–e815. https://doi.org/10.1016/S2589-7500(22)00169-8.
7. Dumitrascu, O. M.; Li, X.; Zhu, W.; Woodruff, B. K.; Nikolova, S.; Sobczak, J.; Youssef, A.; Saxena, S.; Andreev, J.; Caselli, R. J.; Chen, J. J.; Wang, Y. Color Fundus Photography and Deep Learning Applications in Alzheimer Disease. *Mayo Clin. Proc. Digit. Health* **2024**, *2* (4), 548–558. https://doi.org/10.1016/j.mcpdig.2024.08.005.
8. Li, Z.; Wu, X.; Du, H.; Liu, F.; Nghiem, H.; Shi, G. A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; 2025; pp 1578–1597. https://doi.org/10.1109/CVPRW67362.2025.00147.
9. Chen, R. J.; Ding, T.; Lu, M. Y.; Williamson, D. F. K.; Jaume, G.; Song, A. H.; Chen, B.; Zhang, A.; Shao, D.; Shaban, M.; Williams, M.; Oldenburg, L.; Weishaupt, L. L.; Wang, J. J.; Vaidya, A.; Le, L. P.; Gerber, G.; Sahai, S.; Williams, W.; Mahmood, F. Towards a General-Purpose Foundation Model for Computational Pathology. *Nat. Med*. **2024**, *30* (3), 850–862. https://doi.org/10.1038/s41591-024-02857-3.
10. Lu, M. Y.; Chen, B.; Williamson, D. F. K.; Chen, R. J.; Liang, I.; Ding, T.; Jaume, G.; Odintsov, I.; Le, L. P.; Gerber, G.; Parwani, A. V.; Zhang, A.; Mahmood, F. A Visual-Language Foundation Model for Computational Pathology. *Nat. Med.* **2024**, *30* (3), 863–874. https://doi.org/10.1038/s41591-024-02856-4.
11. Tanno, R.; Barrett, D. G. T.; Sellergren, A.; Ghaisas, S.; Dathathri, S.; See, A.; Welbl, J.; Lau, C.; Tu, T.; Azizi, S.; Singhal, K.; Schaekermann, M.; May, R.; Lee, R.; Man, S.; Mahdavi, S.; Ahmed, Z.; Matias, Y.; Barral, J.; Eslami, S. M. A.; Belgrave, D.; Liu, Y.; Kalidindi, S. R.; Shetty, S.; Natarajan, V.; Kohli, P.; Huang, P.-S.; Karthikesalingam, A.; Ktena, I. Collaboration between Clinicians and Vision–Language Models in Radiology Report Generation. *Nat. Med*. **2025**, *31* (2), 599–608. https://doi.org/10.1038/s41591-024-03302-1.
12. Zhou, Y.; Chia, M. A.; Wagner, S. K.; Ayhan, M. S.; Williamson, D. J.; Struyven, R. R.; Liu, T.; Xu, M.; Lozano, M. G.; Woodward-Court, P.; Kihara, Y.; Altmann, A.; Lee, A. Y.; Topol, E. J.; Denniston, A. K.; Alexander, D. C.; Keane, P. A. A Foundation Model for Generalizable Disease Detection from Retinal Images. *Nature* **2023**, *622* (7981), 156–163. https://doi.org/10.1038/s41586-023-06555-x.
13. Wu, Y.; Qian, B.; Li, T.; Qin, Y.; Guan, Z.; Chen, T.; Jia, Y.; Zhang, P.; Zeng, D.; Moroi, S.; Raman, R.; Thinggaard, B. S.; Pedersen, F.; Ñehe, J. A. O.; Kamalden, T. A.; Zhou, Y.; Jin, Y.; Li, H.; Ran, A. R.; Yang, D.; Meng, Z.; Peng, Q.; Zheng, Y. F.; Wang, D.; Ji, H.; Zang, P.; Yin, C.; Shen, J.; Chen, Y.; Yu, W.; Dai, R.; Zhang, C.; Zhao, X.; Wang, X.; Chen, Y.; Wu, Q.; Xie, H.; Szeto, S. K. H.; Chan, J. Y. Y.; Chan, V. T. T.; Xie, H.-T.; Wei, R.; Li, J.; Ma, W.; Zhu, L.; Wang, H.; Fu, H.; Wang, W.; Lin, S.; Xu, Z.; Guan, N.; Zhang, X.; Grzybowski, A.; Gołębiowska-Bogaj, M.; Gawęcki, M.; Smedowski, A.; Szaraniec, W.; Wu, Y.; Wen, Y.; Chen, X.; Yao, Y.; Lim, L.-L.; Cheung, C. Y.; Tan, G. S. W.; Grauslund, J.; Ruamviboonsuk, P.; Sivaprasad, S.; Keane, P. A.; Bian, X.; Wang, Y. X.; Tham, Y.-C.; Cheng, C.-Y.; Wong, T. Y.; Sheng, B. An Eye-care Foundation Model for Clinical Assistance: A Randomized Controlled Trial. *Nat. Med*. **2025**, 1–10. https://doi.org/10.1038/s41591-025-03900-7.
14. Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; Arx, S. von; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; Brynjolfsson, E.; Buch, S.; Card, D.; Castellon, R.; Chatterji, N.; Chen, A.; Creel, K.; Davis, J. Q.; Demszky, D.; Donahue, C.; Doumbouya, M.; Durmus, E.; Ermon, S.; Etchemendy, J.; Ethayarajh, K.; Fei-Fei, L.; Finn, C.; Gale, T.; Gillespie, L.; Goel, K.; Goodman, N.; Grossman, S.; Guha, N.; Hashimoto, T.; Henderson, P.; Hewitt, J.; Ho, D. E.; Hong, J.; Hsu, K.; Huang, J.; Icard, T.; Jain, S.; Jurafsky, D.; Kalluri, P.; Karamcheti, S.; Keeling, G.; Khani, F.; Khattab, O.; Koh, P. W.; Krass, M.; Krishna, R.; Kuditipudi, R.; Kumar, A.; Ladhak, F.; Lee, M.; Lee, T.; Leskovec, J.; Levent, I.; Li, X. L.; Li, X.; Ma, T.; Malik, A.; Manning, C. D.; Mirchandani, S.; Mitchell, E.; Munyikwa, Z.; Nair, S.; Narayan, A.; Narayanan, D.; Newman, B.; Nie, A.; Niebles, J. C.; Nilforoshan, H.; Nyarko, J.; Ogut, G.; Orr, L.; Papadimitriou, I.; Park, J. S.; Piech, C.; Portelance, E.; Potts, C.; Raghunathan, A.; Reich, R.; Ren, H.; Rong, F.; Roohani, Y.; Ruiz, C.; Ryan, J.; Ré, C.; Sadigh, D.; Sagawa, S.; Santhanam, K.; Shih, A.; Srinivasan, K.; Tamkin, A.; Taori, R.; Thomas, A. W.; Tramèr, F.; Wang, R. E.; Wang, W.; Wu, B.; Wu, J.; Wu, Y.; Xie, S. M.; Yasunaga, M.; You, J.; Zaharia, M.; Zhang, M.; Zhang, T.; Zhang, X.; Zhang, Y.; Zheng, L.; Zhou, K.; Liang, P. On the Opportunities and Risks of Foundation Models. arXiv July 12, 2022. https://doi.org/10.48550/arXiv.2108.07258.
15. Topol, E. J. High-Performance Medicine: The Convergence of Human and Artificial Intelligence. *Nat. Med.* **2019**, *25* (1), 44–56. https://doi.org/10.1038/s41591-018-0300-7.
16. Bycroft, C.; Freeman, C.; Petkova, D.; Band, G.; Elliott, L. T.; Sharp, K.; Motyer, A.; Vukcevic, D.; Delaneau, O.; O'Connell, J.; Cortes, A.; Welsh, S.; Young, A.; Effingham, M.; McVean, G.; Leslie, S.; Allen, N.; Donnelly, P.; Marchini, J. The UK Biobank Resource with Deep Phenotyping and Genomic Data. *Nature* **2018**, *562* (7726), 203–209. https://doi.org/10.1038/s41586-018-0579-z.
17. Tian, J.; Smith, G.; Guo, H.; Liu, B.; Pan, Z.; Wang, Z.; Xiong, S.; Fang, R. Modular Machine Learning for Alzheimer's Disease Classification from Retinal Vasculature. *Sci. Rep*. **2021**, *11* (1), 238. https://doi.org/10.1038/s41598-020-80312-2.
18. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodri-

DOI: 10.36838/IJHSR812.62

guez, A.; Joulin, A.; Grave, E.; Lample, G. LLaMA: Open and Efficient Foundation Language Models. arXiv February 27, 2023. https://doi.org/10.48550/arXiv.2302.13971.
19. Liu, H.; Li, C.; Wu, Q.; Lee, Y. J. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Curran Associates, Inc., 2023; Vol. 36, pp 34892–34916.
20. Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; Gao, J. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*; NIPS '23; Curran Associates Inc.: Red Hook, NY, USA, 2023.
21. Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; Zhou, J. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *ArXiv Prepr. ArXiv230812966* **2023**.
22. Zhu, W.; Li, X.; Chen, X.; Qiu, P.; Vasa, V. K.; Dong, X.; Chen, Y.; Lepore, N.; Dumitrascu, O.; Su, Y.; Wang, Y. RetinalGPT: A Retinal Clinical Preference Conversational Assistant Powered by Large Vision-Language Models. arXiv March 6, 2025. https://doi.org/10.48550/arXiv.2503.03987.
23. Sánchez, C. I.; Niemeijer, M.; Dumitrescu, A. V.; Suttorp-Schulten, M. S. A.; Abràmoff, M. D.; van Ginneken, B. Evaluation of a Computer-Aided Diagnosis System for Diabetic Retinopathy Screening on Public Data. *Invest. Ophthalmol. Vis. Sci.* **2011**, *52* (7), 4866–4871. https://doi.org/10.1167/iovs.10-6633.
24. *APTOS 2019 Blindness Detection*. https://kaggle.com/competitions/aptos2019-blindness-detection (accessed 2023-03-24).
25. Fu, H.; Wang, B.; Shen, J.; Cui, S.; Xu, Y.; Liu, J.; Shao, L. Evaluation of Retinal Image Quality Assessment Networks in Different Color-Spaces. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*; Shen, D., Liu, T., Peters, T. M., Staib, L. H., Essert, C., Zhou, S., Yap, P.-T., Khan, A., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2019; pp 48–56. https://doi.org/10.1007/978-3-030-32239-7_6.
26. Porwal, P.; Pachade, S.; Kamble, R.; Kokare, M.; Deshmukh, G.; Sahasrabuddhe, V.; Meriaudeau, F. Indian Diabetic Retinopathy Image Dataset (IDRiD): A Database for Diabetic Retinopathy Screening Research. *Data* **2018**, *3* (3), 25. https://doi.org/10.3390/data3030025.
27. *Myopic Maculopathy Analysis: MICCAI Challenge MMAC 2023, Held in Conjunction with MICCAI 2023, Virtual Event, October 8–12, 2023, Proceedings*; Sheng, B., Chen, H., Wong, T. Y., Eds.; Lecture Notes in Computer Science; Springer Nature Switzerland: Cham, 2024; Vol. 14563. https://doi.org/10.1007/978-3-031-54857-4.
28. Li, N.; Li, T.; Hu, C.; Wang, K.; Kang, H. A Benchmark of Ocular Disease Intelligent Recognition: One Shot for Multi-Disease Detection. arXiv February 16, 2021. https://doi.org/10.48550/arXiv.2102.07978.
29. Pachade, S. Retinal Fundus Multi-Disease Image Dataset (RFMiD), 2020. https://ieee-dataport.org/open-access/retinal-fundus-multi-disease-image-dataset-rfmid (accessed 2023-03-24).
30. Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. de las; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; Sayed, W. E. Mistral 7B. arXiv October 10, 2023. https://doi.org/10.48550/arXiv.2310.06825.
31. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; Kang, J. BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics* **2020**, *36* (4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682.

Authors

Vincent Z Wang is a junior at Brophy College Preparatory in Phoenix, Arizona. He is passionate about medical science and artificial intelligence research. Vincent was a finalist in the 2023 VEX Robotics World Championship – VEX IQ Middle School Event and continues to explore AI applications in healthcare and neuroscience.

Yiyi Sun is a student passionate about the intersection of chemistry and biology. He enjoys using data analysis and computational tools to uncover patterns in scientific data. He plans to study the life sciences in college and hopes to further explore research at the interface of biology, chemistry, and technology.

Dr. Oana M Dumitrascu, M.D., is a Professor of Neurology, vascular neurologist, neuro-ophthalmologist, and medical AI researcher at Mayo Clinic, Arizona. Her work focuses on retinal biomarkers and AI-based applications in stroke and dementia. She is passionate about mentoring and has guided numerous students in neuroscience and medical AI research.