Benchmarking Vision-Language Models for Image-Based Infectious Versus Non-Infectious Uveitis Classification

Yuning Wang[1], Zihao Zhong[1], Yueqin Diao[1], Huiying Liu[2], Jiansheng Fang[3], Kai Li[4], Haoran Zhang[4], Yanwu Xu[1,4⋆], and Rupesh Agrawal[2,5,6⋆⋆]

[1] School of Future Technology, South China University of Technology, Guangzhou 510006, China
202410194185@mail.scut.edu.cn, m13058312522@163.com, ftyueqindiao@mail.scut.edu.cn, yxwu@ieee.org
[2] Department of Ophthalmology, National Healthcare Group Eye Institute, Tan Tock Seng Hospital, Singapore
liuhuiyingcn@gmail.com
[3] Research & Development Institute, Guangzhou Guangri Stock Co., Ltd., Guangzhou 510045, China
11949039@mail.sustech.edu.cn
[4] China Pazhou Lab, Guangzhou 510320, China
leecare10@gmail.com, zhr.hdmi@gmail.com, yxwu@ieee.org
[5] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore
[6] Department of Ophthalmology, National University of Singapore
rupeshttsh@gmail.com

Abstract. Uveitis is a relatively uncommon and etiologically heterogeneous inflammatory eye disease, and infectious-versus-non-infectious classification is clinically important but challenging for direct vision-language model (VLM) classification. Using real-world fundus images from Tan Tock Seng Hospital, Singapore, we built a fixed 438-image benchmark and evaluated five VLM settings under zero-shot generation, LoRA-style adaptation, and frozen-feature linear probing, with image-only encoders as baselines. Zero-shot hard decisions collapsed to a single class across all VLM settings, yielding 0.5000 balanced accuracy, yet several score streams showed non-random AUC. LoRA adaptation was limited and inconsistent, with one MedGemma-27B setting reaching 0.6441 accuracy and 0.6431 balanced accuracy. Frozen-feature probing better converted encoded representations into classification performance, reaching 0.7797 accuracy and 0.7954 AUC for LLaVA-NeXT and 0.7627 accuracy and 0.8391 AUC for InternVL3.5. In selected rows and metrics, these linear-head results were numerically higher than the strongest image-only baseline, DenseNet121, which achieved 0.7458 accuracy and 0.8230 AUC. A secondary threshold analysis of two preserved score streams further suggested recoverable score-level information. These findings reveal a score/representation–decision mismatch in uncommon, fine-grained ophthalmic classification.

⋆ Corresponding author.
⋆⋆ Corresponding author.

Keywords: uveitis · vision-language model · ophthalmology

1 Introduction

Uveitis is a relatively uncommon but vision-threatening inflammatory eye disease with substantial etiologic heterogeneity [8]. Its causes include infectious, non-infectious inflammatory, masquerade, and systemic disease-related conditions, and etiologic misclassification can alter treatment direction and timing [11]. Distinguishing infectious from non-infectious uveitis is therefore a meaningful image-based decision-support task.

Deep learning has been widely used for ophthalmic image classification. ResNet [5] and DenseNet [7] provide strong image-only baselines, while ConvNeXt [10], Vision Transformers [3], and DINOv3 [16] represent modern visual encoders. In parallel, VLMs have extended from natural-scene multimodal understanding to medical and ophthalmic applications. We selected representative settings covering Qwen3-VL, LLaVA-NeXT, InternVL3.5, and MedGemma-family models [2, 9, 19, 15]; MedGemma-4B is multimodal, whereas the 27B setting provides a larger medical language-and-reasoning backbone evaluated through the study-specific image-conditioned interface.

Ophthalmology-oriented benchmarks such as BELO, LMOD, and LMOD+ evaluate ophthalmic knowledge, reasoning, and multimodal understanding [17, 14, 13], but reliable performance on this real-world uveitis classification task remains unclear. More broadly, fine-grained VLM analyses suggest that strong VQA or multimodal reasoning does not necessarily imply reliable fine-grained image classification [4]. We use VLMs not to assume superiority over image-only classifiers, but to examine whether clinically relevant information is retained in their representations and reliably expressed through the generative decision interface. We therefore evaluated five VLM settings under direct zero-shot generation, LoRA-style adaptation, and frozen-feature linear probing, with image-only encoders as baselines. We also examined two preserved score streams to test whether hard-label collapse could mask recoverable score-level signal.

2 Methods

2.1 Dataset and task

This benchmark uses ocular fundus images collected from Tan Tock Seng Hospital, Singapore, for binary infectious-versus-non-infectious uveitis classification. All experiments followed a frozen manifest with image paths, split assignments, case identifiers, and binary labels. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i is an image and $y_i \in \{0, 1\}$, with 0 denoting infectious and 1 denoting non-infectious uveitis. The benchmark contains 438 images: 319 for training, 60 for validation, and 59 for held-out testing, including 30 infectious and 29 non-infectious test images.

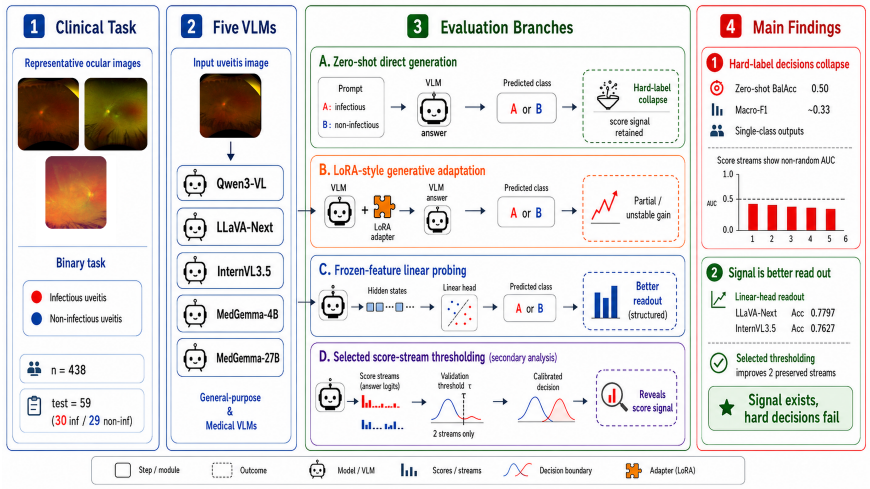

Fig. 1. Benchmark overview. Five VLM settings were evaluated on a fixed real-world uveitis fundus benchmark of 438 images, with a held-out test set of 59 images. The main settings included zero-shot generation, LoRA-style adaptation, frozen-feature probing, and secondary selected score-thresholding.

2.2 VLM evaluation settings

We evaluated five VLM settings: Qwen3-VL [2], LLaVA-NeXT [9], InternVL3.5 [19], MedGemma-4B, and MedGemma-27B [15]. MedGemma-4B denotes the multimodal image-text variant, whereas MedGemma-27B denotes a larger medical language-and-reasoning setting evaluated through the study-specific image-conditioned interface. The main settings were direct zero-shot generation, LoRA-style adaptation, and frozen-feature linear probing. We also trained ResNet50, DenseNet121, ConvNeXt, ViT, and DINOv3 baselines on the same split, and performed targeted threshold analysis only for two preserved score streams.

2.3 Direct generation and LoRA-style adaptation

Zero-shot and LoRA evaluations used the same forced-choice prompt:

```
<image>
Classify this ocular image for binary uveitis classification.
A = infectious uveitis
B = non-infectious uveitis
Output only one letter: A or B.
```

Generated answers were normalized to binary labels by parsing the final standalone token as either A or B. LoRA-style adaptation used supervised fine-tuning [6] on image-prompt-answer triples with the same answer format. Across the reported runs, the adapters used rank 16, alpha 32, dropout 0.05, learning rate 2×10^{-4}, five epochs, per-device batch size 1, and gradient accumulation 8. Most models used all supported linear modules as LoRA targets. For

InternVL3.5, LoRA was restricted to attention and MLP projection layers to reduce memory cost and avoid destabilizing its larger multimodal architecture.

2.4 Frozen-feature linear probing

For frozen-feature probing, we tested whether diagnostic information was encoded in each evaluated model setting independently of its generative decision head. Following a linear-probing protocol [1], each feature source was kept frozen. For each image, we used the prompt "`<image>\nDescribe briefly.`" while requesting hidden states; no generated text was used for classification. The final-layer hidden-state sequence was mean-pooled into one fixed-length representation per image without additional post-pooling normalization.

A model-specific linear classifier was then trained on top of each frozen representation space for infectious-versus-non-infectious prediction. This design isolates the diagnostic utility of the pretrained representation from the behavior of the autoregressive output interface. Linear heads were optimized on the training split using AdamW with a learning rate of 1×10^{-3}, weight decay of 1×10^{-2}, batch size 64, and a maximum of 100 epochs. Class-balanced cross-entropy was used to account for training-set class imbalance. Validation macro-F1 was used for early stopping and checkpoint selection with a patience of 10 epochs, after which the selected checkpoint was evaluated once on the held-out test split. Softmax outputs from the linear head were used both for hard-label prediction and for continuous score-based metrics.

2.5 Targeted threshold analysis of preserved score streams

This secondary analysis was restricted to two preserved continuous score streams with validation and test scores. It was not a full-model calibration benchmark, but a targeted test of whether collapsed hard predictions could retain recoverable score information, motivated by VLM calibration and threshold-sensitive evaluation work [18, 12]. For each stream, the score was the preserved non-infectious-side probability $s_i = p(\text{non-infectious} \mid x_i, q)$. Hard decisions were defined as $\hat{y}_i(\tau) = \mathbb{1}[s_i \geq \tau]$, with τ selected on validation predictions by maximizing macro-F1 and then evaluated on the held-out test predictions.

2.6 Image-only baselines and metrics

Image-only baselines used resized 224×224 inputs, standard augmentation, AdamW, 20 epochs, batch size 32, learning rate 10^{-4}, weight decay 10^{-4}, and validation AUC model selection. These baselines are contextual comparators rather than matched replacements for the VLM settings. Main metrics were accuracy, balanced accuracy, macro-F1, AUC when a continuous score was available, and class-specific infectious and non-infectious recall.

3 Results

3.1 Direct zero-shot generation

Direct zero-shot generation produced degenerate hard-label behavior: each VLM setting collapsed to a single predicted class on the held-out test set (Table 1). Qwen3-VL and InternVL3.5 predicted 59 infectious and 0 non-infectious cases, whereas LLaVA-NeXT and both MedGemma settings predicted 0 infectious and 59 non-infectious cases. In both situations, the class recalls were 1.0000 and 0.0000, yielding a balanced accuracy of $(1.0000 + 0.0000)/2 = 0.5000$. AUC varied from 0.3431 to 0.6207, with several streams exceeding chance and one showing an inverse ranking tendency. Thus, hard-label collapse did not imply absence of score-level signal; rather, the default generative interface failed to convert available ranking information into balanced predictions.

3.2 LoRA-style adaptation

LoRA-style adaptation partially changed this behavior but did not consistently stabilize direct generation (Table 1). Most adapted settings still produced poor or imbalanced hard-label decisions. Qwen3-VL improved modestly to 0.5424 accuracy and 0.5402 balanced accuracy, whereas one MedGemma-27B setting improved to 0.6441 accuracy, 0.6431 balanced accuracy, and 0.6810 AUC. These results suggest that lightweight generative adaptation can sometimes improve the decision interface, but does not reliably resolve the mismatch between score-level signal and final generated labels.

Table 1. Direct zero-shot and LoRA-style generative results on the held-out test set ($n = 59$). A balanced accuracy of 0.5000 indicates single-class prediction, with class recalls of 1.0000 and 0.0000.

Setting	Model	Acc	AUC	BalAcc	Macro-F1	Inf. Rec.	Non-inf. Rec.
Zero-shot	Qwen3-VL	0.5085	0.3431	0.5000	0.3371	1.0000	0.0000
	LLaVA-NeXT	0.4915	0.5425	0.5000	0.3295	0.0000	1.0000
	InternVL3.5	0.5085	0.5000	0.5000	0.3371	1.0000	0.0000
	MedGemma-4B	0.4915	0.6207	0.5000	0.3295	0.0000	1.0000
	MedGemma-27B	0.4915	0.5943	0.5000	0.3295	0.0000	1.0000
LoRA	Qwen3-VL	0.5424	0.5259	0.5402	0.5338	0.6667	0.4138
	LLaVA-NeXT	0.5085	0.5230	0.5000	0.3371	1.0000	0.0000
	InternVL3.5	0.5085	0.5000	0.5000	0.3371	1.0000	0.0000
	MedGemma-4B	0.5085	0.6511	0.5000	0.3371	1.0000	0.0000
	MedGemma-27B	0.6441	0.6810	0.6431	0.6424	0.7000	0.5862

3.3 Frozen-feature linear probing

Frozen-feature probing provided a more structured readout of the diagnostic information encoded by the pretrained VLMs (Table 2), with the zero-shot-to-probing performance shift summarized in Fig. 2. Unlike direct generation, the

Table 2. Frozen-feature linear probing results ($n = 59$).

Model	Acc	AUC	BalAcc	Macro-F1	Inf. Rec.	Non-inf. Rec.
Qwen3-VL	0.5932	0.8414	0.5862	0.5042	1.0000	0.1724
LLaVA-NeXT	0.7797	0.7954	0.7776	0.7755	0.9000	0.6552
InternVL3.5	0.7627	0.8391	0.7644	0.7610	0.6667	0.8621
MedGemma-4B	0.7288	0.7701	0.7293	0.7287	0.7000	0.7586
MedGemma-27B	0.7627	0.8023	0.7609	0.7593	0.8667	0.6552

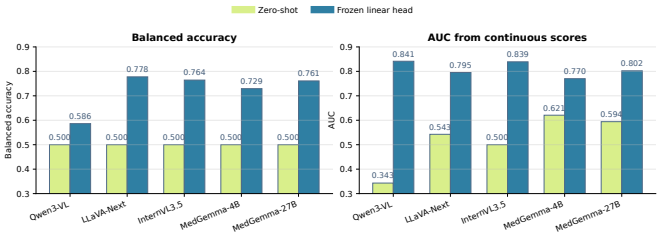

Fig. 2. Zero-shot generation versus frozen-feature probing.

linear heads were trained to map frozen representations directly to the binary label space, thereby separating representation quality from the autoregressive answer interface. This setting yielded stronger and more balanced classification performance, with LLaVA-NeXT reaching 0.7797 accuracy and 0.7954 AUC, InternVL3.5 reaching 0.7627 accuracy and 0.8391 AUC, and MedGemma-27B reaching 0.7627 accuracy and 0.8023 AUC. Within the linear-probing setting, Qwen3-VL showed the clearest ranking–decision gap: its probe achieved the highest AUC in the table (0.8414), but its thresholded hard-label performance remained weaker at 0.5932 accuracy and 0.5862 balanced accuracy. These gains indicate that clinically relevant information was present in the frozen representations and could be better exploited by a lightweight discriminative readout.

3.4 Targeted threshold analysis

Validation-selected thresholding on two preserved score streams further supported this interpretation (Table 3). InternVL3.5 improved from 0.5000 to 0.7138 balanced accuracy at the same AUC of 0.7471, and Qwen3-VL improved from 0.5000 to 0.6966 balanced accuracy with AUC 0.7241. This analysis was limited to two streams and should be interpreted as targeted evidence for recoverable score signal rather than a full-model calibration benchmark.

3.5 Image-only visual baselines

Image-only baselines provided additional evidence of recoverable visual signal (Table 4). DenseNet121 was strongest, with 0.7458 accuracy, 0.7443 balanced accuracy, and 0.8230 AUC. ResNet50 and DINOv3 were intermediate, ViT was weaker, and ConvNeXt collapsed to all-infectious predictions. In selected rows,

Table 3. Targeted threshold analysis of two preserved score streams ($n = 59$).

Model	Decision rule	Acc	AUC	BalAcc	Macro-F1	Inf. Rec.	Non-inf. Rec.
InternVL3.5	Def. ($\tau = 0.500$)	0.5085	0.7471	0.5000	0.3371	1.0000	0.0000
InternVL3.5	Val sel. ($\tau = 0.212$)	0.7119	0.7471	0.7138	0.7089	0.6000	0.8276
Qwen3-VL	Def. ($\tau = 0.500$)	0.5085	0.7241	0.5000	0.3371	1.0000	0.0000
Qwen3-VL	Val sel. ($\tau = 0.301$)	0.6949	0.7241	0.6966	0.6927	0.6000	0.7931

the linear-head results of InternVL3.5 and LLaVA-NeXT were numerically higher than the strongest image-only baseline in accuracy, while InternVL3.5 was also numerically higher in AUC. These cross-family comparisons are descriptive because the training settings are not fully matched.

Table 4. Image-only visual encoder baselines ($n = 59$).

Model	Acc	AUC	BalAcc	Macro-F1	Inf. Rec.	Non-inf. Rec.
ResNet50	0.6271	0.6839	0.6218	0.5840	0.9333	0.3103
DenseNet121	0.7458	0.8230	0.7443	0.7431	0.8333	0.6552
ConvNeXt	0.5085	0.5172	0.5000	0.3371	1.0000	0.0000
ViT	0.5085	0.6046	0.5126	0.4791	0.2667	0.7586
DINOv3	0.6271	0.7069	0.6264	0.6262	0.6667	0.5862

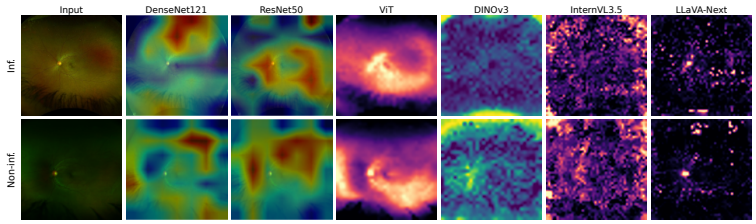

Fig. 3. Representative qualitative maps. CNN maps are Grad-CAM overlays; transformer and VLM maps are token-level visualizations. These maps are qualitative only.

4 Discussion

The main finding is a score/representation–decision mismatch rather than a complete absence of diagnostic signal. Zero-shot outputs collapsed to single-class predictions, yet several score streams showed non-random AUC, because AUC evaluates continuous-score ranking whereas accuracy depends on the final thresholded or generated label. Thus, available score-level information can be poorly expressed by the default generative interface.

LoRA-style adaptation provided only partial and inconsistent improvement, whereas frozen-feature probing more effectively converted encoded representations into balanced predictions. The selected score-stream analysis supports this interpretation but remains limited to two streams and should not be viewed as a general calibration strategy. Likewise, frozen-feature probing is a diagnostic readout rather than a deployment strategy. Overall, these findings support evaluating medical VLMs across generated answers, continuous scores, and internal representations. This study remains an exploratory single-center benchmark with

a modest 59-image test set and no external validation or uncertainty estimates. Numerical differences between model families should therefore be interpreted descriptively, particularly because the frozen probes and image-only baselines are not fully matched. Future work should include repeated patient-level evaluation and external validation across hospitals and imaging systems.

5 Conclusion

In this real-world uveitis benchmark, direct VLM generation was unreliable as a hard-label decision interface, whereas score-level behavior and frozen-feature probing indicated useful diagnostic signal in selected score streams and frozen representations. Lightweight adaptation only partially improved hard decisions, while discriminative readouts better used encoded signal, highlighting the need to evaluate both what medical VLMs encode and whether their decision interface can reliably express it.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 62571192).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alain, G., Bengio, Y.: Understanding intermediate layers using linear classifier probes (2016). https://doi.org/10.48550/arXiv.1610.01644
2. Bai, S., Cai, Y., Chen, R., et al.: Qwen3-vl technical report (2025). https://doi.org/10.48550/arXiv.2511.21631
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy
4. Ghosh, D., Zhang, Y., Schmidt, L.: Understanding the fine-grained knowledge capabilities of vision-language models (2026). https://doi.org/10.48550/arXiv.2602.17871
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
7. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4700–4708 (2017)

8. Jacquot, R., Ren, L., Wang, T., Mellahk, I., Duclos, A., Kodjikian, L., Jamilloux, Y., Stanescu, D., Sève, P.: Neural networks for predicting etiological diagnosis of uveitis. Eye **39**, 992–1002 (2025). https://doi.org/10.1038/s41433-024-03530-2
9. Li, B., Zhang, K., Zhang, H., et al.: LLaVA-NeXT: Stronger LLMs Supercharge Multimodal Capabilities in the Wild (2024), `https://llava-vl.github.io/blog/2024-01-30-llava-next/`
10. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986 (2022)
11. Maghsoudlou, P., Epps, S.J., Guly, C.M., Dick, A.D.: Uveitis in adults: A review. JAMA **334**(5), 419–434 (2025). https://doi.org/10.1001/jama.2025.4358
12. Oh, C., Lim, H., Kim, M., Han, D., Yun, S., Choo, J., Hauptmann, A., Cheng, Z.Q., Song, K.: Towards calibrated robust fine-tuning of vision-language models (2023). https://doi.org/10.48550/arXiv.2311.01723
13. Qin, Z., Liu, Y., Yin, Y., Ding, J., Zhang, H., et al.: Lmod+: A comprehensive multimodal dataset and benchmark for developing and evaluating multimodal large language models in ophthalmology (2025). https://doi.org/10.48550/arXiv.2509.25620
14. Qin, Z., Yin, Y., Campbell, D., Wu, X., Zou, K., Tham, Y.C., Liu, N., Zhang, X., Chen, Q.: Lmod: A large multimodal ophthalmology dataset and benchmark for large vision-language models (2024). https://doi.org/10.48550/arXiv.2410.01620
15. Sellergren, A., Kazemzadeh, S., Jaroensri, T., et al.: Medgemma technical report (2025). https://doi.org/10.48550/arXiv.2507.05201
16. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., et al.: Dinov3 (2025). https://doi.org/10.48550/arXiv.2508.10104
17. Srinivasan, S., Ai, X., Lo, T.W.S., et al.: Benchmarking llms for ophthalmology (belo) for ophthalmological knowledge and reasoning (2025). https://doi.org/10.48550/arXiv.2507.15717
18. Tu, W., Deng, W., Campbell, D., Gould, S., Gedeon, T.: An empirical study into what matters for calibrating vision-language models (2024). https://doi.org/10.48550/arXiv.2402.07417
19. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., et al.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency (2025). https://doi.org/10.48550/arXiv.2508.18265