IRIS: An Intelligent Vision-Language System for Ocular Surface Diseases via Topic Tree and Scene-Driven VQA Generation

Hao Wei[1⋆][0000−0002−5719−8826], Wenjin Qi[2⋆][0009−0004−9816−1423], Dasen Dai[1][0009−0000−3227−3140], Minqing Zhang[1][0000−0002−7214−0569], and Wu Yuan[1][0000−0001−9405−519X]⋆⋆

[1] The Chinese University of Hong Kong, Shatin, Hong Kong, China
[2] Westlake University, Xihu District, Hangzhou, Zhejiang Province, China

Abstract. While Large Vision-Language Models (VLMs) demonstrate remarkable generic capabilities, their clinical reasoning in specialized domains like ocular surface diseases (OSDs) is severely hindered by a paucity of high-fidelity, multimodal instruction-tuning data. To dismantle this data bottleneck, we introduce IRIS, an Intelligent Recognition and Interaction System tailored for fine-grained OSD understanding via external eye photography. First, we curate IRIS-120K, the largest and most comprehensive OSD visual question-answering (VQA) dataset to date. Crucially, to overcome the semantic shallowness of conventional image-caption pairs, we propose a synergistic data generation paradigm to explicitly inject clinical priors. Our data engine operates via a dual-branch framework: 1) a Topic Finding Tree (TFT) that hierarchically anchors visual features to precise anatomical and pathological concepts, enforcing rigorous medical deduction logic; and 2) a Scene-driven strategy that synthesizes role-adaptive clinical dialogues to ensure pragmatic generalization. By explicitly aligning a compact 4B-parameter VLM on this structurally enriched corpus, IRIS achieves highly competitive performance and outperforms both generalist and specialized medical VLMs with up to 34B parameters within this dataset. Our findings underscore that structured knowledge injection effectively bridges the domain gap, unlocking the potential for resource-efficient, expert-level AI deployment on mobile edge devices for scalable OSD screening. Code, datasets, and model weights will be publicly released by this repo.

Keywords: Vision-Language Models · Ocular Surface Diseases · Clinical Reasoning · Interpretability · Mobile Edge AI.

1 Introduction

Ocular surface diseases (OSDs), including dry eye disease (DED) and infectious keratitis, represent a formidable global health challenge and the fifth leading

⋆ Equal contribution.
⋆⋆ Corresponding author (wyuan@cuhk.edu.hk).

cause of blindness [15,26]. DED afflicts up to 34% of adults, while infectious keratitis is a leading cause of blindness, which is largely preventable if diagnosed early [6,15]. Despite this prevalence, expert-level diagnosis traditionally mandates specialized stationary equipment (e.g., slit-lamp biomicroscopes) [13,26], exacerbating healthcare disparities in resource-limited settings. Consequently, the ubiquity of high-resolution smartphone cameras has catalyzed mobile eye health (MeHealth) into a transformative frontier, offering an unprecedented opportunity for accessible, large-scale OSD screening via external eye photography [23].

Ophthalmic AI has rapidly evolved from task-specific classifiers (e.g., CorneAI [14,15]) towards multimodal foundation models like RETFound [28], VisionFM [17] and EyeCLIP [19]. While these architectures demonstrate robust generalization for retinal conditions, a critical modality disparity persists: current models exhibit a profound bias towards the posterior segment. For instance, within the 2.77-million-image EyeCLIP dataset, external eye photographs account for a negligible 0.5%. Similarly, EyecareGPT [9] predominantly targets fundus and OCT imagery. This leaves a severe data and reasoning vacuum in the fine-grained understanding of external eye pathologies.

While generic Large Vision-Language Models (VLMs) show potential for interactive diagnostics [21,25], recent benchmarking exposes their stark clinical limitations in ophthalmology, often producing hallucinated interpretations lacking precise anatomical grounding [25,10]. Even domain-specific medical VLMs (e.g., HuatuoGPT-Vision [4], Lingshu [24]) are inherently bottlenecked by the flat, unstructured nature of PubMed image-caption pairs. Consequently, they consistently fail to spatially anchor morphological features to specific anatomical structures (e.g., distinguishing a corneal infiltrate from an adjacent iris nodule)—a rigorous spatial deduction step critical for accurate OSD diagnosis [8].

To bridge this diagnostic gap, we introduce **IRIS** (**I**ntelligent **R**ecognition & **I**nteraction **S**ystem for OSDs). We curate **IRIS-120K**, the largest comprehensive OSD visual question-answering (VQA) dataset to date. To explicitly overcome the semantic misalignment of generic VLMs [8,25], we engineer a Clinically-driven Data Engine composed of two novel paradigms: 1) **Topic Finding Tree (TFT):** A structured framework that hierarchically anchors visual features to 10 predefined ocular regions and pathological findings, forcing the model to internalize rigorous, step-by-step clinical deduction. 2) **Scene-Driven Generation:** A role-adaptive paradigm simulating clinical interactions across three primary users (Doctor, Patient, Student) and 12 scenarios, ensuring pragmatic utility in real-world MeHealth deployments.

By fine-tuning a highly compact 4B-parameter VLM on this enriched corpus, IRIS demonstrates strong performance on our curated dataset, overwhelmingly surpassing generalist and medical models possessing up to 34B parameters. Our findings underscore a vital paradigm shift: meticulous structural knowledge injection and explicit anatomical grounding—rather than mere parameter scaling—are effective strategies to enabling highly efficient, expert-level AI deployment on mobile edge devices for scalable OSD screening [8,23].

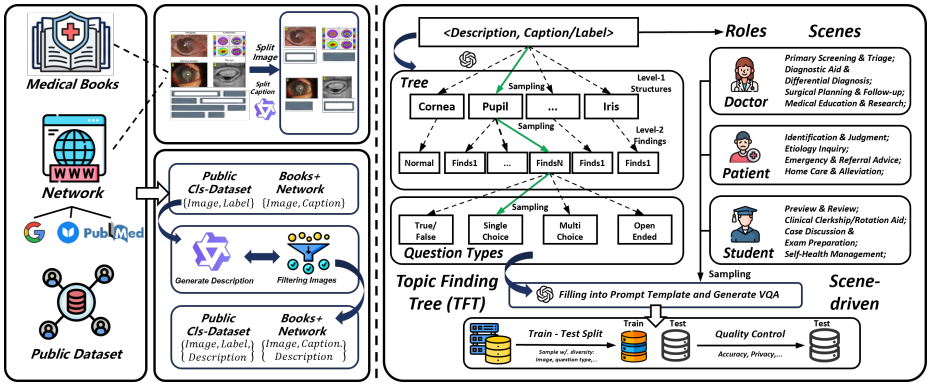

Fig. 1. Overview of our Clinically-driven Data Engine. The pipeline aggregates heterogeneous ocular images, and utilizes a dual-branch VQA generation paradigm: Topic Finding Tree (TFT) for rigorous clinical reasoning and Scene-driven generation for role-adaptive interactions.

The primary contributions of this work are summarized as follows:

- **A Dedicated OSD Interactive Corpus:** We curate and release IRIS-120K, successfully filling a major void in high-fidelity, multimodal instruction-tuning data specifically tailored for external eye photography.
- **Knowledge-Injected Data Engines:** We pioneer the TFT and Scene-Driven strategies to explicitly model spatial ocular anatomy and clinical personas, fundamentally resolving the interpretability and pragmatism deficits inherent in existing medical VLMs.
- **Highly Efficient Foundation Model:** We open-source IRIS-4B, a lightweight clinical VLM optimized for resource-constrained environments. It demonstrates unprecedented parameter efficiency and fine-grained visual-textual grounding compared to massively scaled baselines.

2 Methodology

To bridge the gap between the poor zero-shot performance of existing VLMs and the clinical demand for OSD screening, we propose a novel Clinically-driven Data Engine (Fig. 1). Rather than simply scaling up model parameters, our framework curates **IRIS-120K**, the largest external eye VQA dataset to date. By innovating Topic Finding Tree (TFT) and Scene-driven generation paradigms, we inject rigorous medical reasoning and role-adaptive empathy into lightweight models, enabling highly efficient edge deployment on patient smartphones.

2.1 Large-Scale Data Curation and Preprocessing

To construct our Data Engine, we aggregate external eye images into *image-caption* or *image-label* pairs from three heterogeneous sources: **Medical Books**

(3.1K cases), **Network Data** (93.5K samples from WeChat, PubMed, and Google Lens), and **Public Datasets** (35.4K). Note that these counts are the results before any filtering. To resolve domain noise and format inconsistencies, we design source-specific preprocessing pipelines.

Books and PubMed Literature. Medical literature contains high-value but complex multi-panel figures that severely degrade VLM image-text alignment. Sourcing from textbooks and BIOMEDICA [12] (a biomedical image-caption archive), we deploy a rigorous three-stage pipeline [1]. First, Qwen2.5-VL-7B [3] filters non-external eye images and categorizes data into Paper-Single (standalone) and Paper-Multi (compound). For *Paper-Multi*, a YOLOv7-based detector [22] segments sub-figures, and Qwen2.5-VL-72B [3] uses these bounding boxes to semantically map the overarching monolithic caption to specific subfigure crops. A secondary VLM filter eliminates irrelevant content (e.g., fundus) exposed after splitting. Captions are further enriched using main-text references. Book data undergoes an identical fine-grained alignment.

Web Resources. To capture real-world long-tail diversity, we use curated book images as visual seeds for Google Lens reverse searches, extracting alt-texts as captions. Concurrently, we scrape ophthalmic WeChat articles, which natively provide structured *(image, disease name, symptom description)* triplets. Both sources undergo strict VLM-based filtering to guarantee external-eye relevance.

Public Datasets Adaptation. We integrate high-quality classification and segmentation datasets. To adapt segmentation data, we introduce an area-based heuristic: if over 50% of the images contain lesion bounding boxes occupying $\geq$50% of the total image area (meaning localized pathology visually dominates), the entire dataset is repurposed for classification. All data is standardized as *(image, label)* pairs.

Quality-Aware Stratification. Given varying raw source fidelity, we conduct sampling-based human evaluations on resolution, image-text alignment, and semantic richness to derive a strict quality hierarchy: **WeChat** > **Books** > **Paper-Single** > **Public Cls** > **Paper-Multi** > **Web (Google Lens)**. This empirically established prior fundamentally directs our "Quality-Aware Dynamic Sampling" strategy (Sec. 2.2), ensuring high-fidelity data dominates the training distribution of our lightweight models.

2.2 Knowledge-Driven VQA Generation & Quality-Aware Sampling

Traditional "image-caption" pairs generate flat representations, failing to impart the "anatomy-to-sign" deduction logic required in clinical practice. To overcome this, we transform the curated images into interpretable VQA pairs using a dual-branch strategy. To enrich the semantic context and mitigate LLM hallucinations, we first employ Qwen3-VL-32B [2] to generate a detailed visual description T_{desc} by the image and its original source information, including captions, disease labels, symptom descriptions, or textbook/public-dataset annotations. Combined with the original caption/label T_{cap}, this forms a rich textual surrogate $x = (T_{cap}, T_{desc})$, serving as the foundational semantic anchor for downstream generation.

Topic Finding Tree (TFT) Generation. To force the model to "think step-by-step like an ophthalmologist", we introduce TFT, a structured anatomical decomposition framework. The TFT hierarchically parses x into a depth-2 logical tree: the first level comprises 10 predefined ocular anatomical regions $a_i \in \mathcal{A}$ (*Pupil, Cornea, Iris, Sclera-Conjunctiva Region, Eyelashes, Upper Eyelid, Lower Eyelid, Superior Palpebral Conjunctiva, Inferior Palpebral Conjunctiva, and Others*), and the second level extracts region-specific clinical findings $f \in F_{a_i}$. This explicitly establishes anatomical-pathological paths $\mathcal{T} = \{(a_i, f) \mid a_i \in \mathcal{A}, f \in F_{a_i}\}$.

For each instantiated path, an LLM generates diverse question formats (e.g., True/False, Single-choice, Multiple-choice or Open-ended). Crucially, we prompt the LLM to construct a structured **4-stage reasoning chain** (*Think*): *(1) Visual Observation* (identifying morphological features); *(2) Clinical Correlation*; *(3) Logical Deduction* (excluding differential diagnoses); and *(4) Conclusion*. This explicit mapping ensures the generated VQA tuples $(Q, A, Think)$ are strictly grounded in localized clinical evidence.

Scene-Driven Adaptation. To accommodate the diverse interaction needs of real-world edge deployments, we simulate conversational contexts across three primary user roles: $\mathcal{R} = \{\text{Patient, Doctor, Student}\}$. For each role, we define tailored clinical scenes $\mathcal{S}_r$ (e.g., triage advice for patients, differential diagnosis for doctors, as shown in the right panel of the Fig. 1). Using scene-specific templates, the LLM generates interactions incorporating a sequential clinical chain-of-thought—*Query Analysis, Evidence Extraction, Clinical Reasoning, Response Formulation*—ensuring the response tone and jargon density closely align the target user's medical literacy.

Quality-Aware Dynamic Sampling & Dataset Finalization. Exhaustive generation across all TFT nodes and clinical scenes would yield over 2 million highly redundant VQA pairs. This contradicts our fundamental objective of training a highly efficient, small-parameter model (2B/4B) capable of offline edge deployment. Therefore, we implement a quality-aware dynamic sampling protocol. Instead of uniform generation, we probabilistically assign combination strategies based on the predefined data quality hierarchy. High-quality sources trigger denser generation (e.g., sampling multiple TFT regions and scenarios simultaneously), while lower-quality web data is sparsely sampled to maximize information yield without redundancy. For each image, 1–3 prompts are generated using either the TFT or Scene-Driven method.

This balanced protocol generates approximately 130K initial VQA pairs. We apply pHash [27] for strict image deduplication to prevent data leakage. An initial test pool of 9K VQA pairs undergoes a rigorous final quality control pass using GPT-5-mini to filter out factual hallucinations and privacy anomalies. To further validate data quality, a random subset was reviewed by an ophthalmologist, achieving an 87/100 factual accuracy for image descriptions and a 4.4/5.0 rating for the generated VQA pairs. This ultimately yields the **IRIS-120K**, comprising **117.7K** training pairs and **8.2K** test set (detailed statistics in Fig. 2).

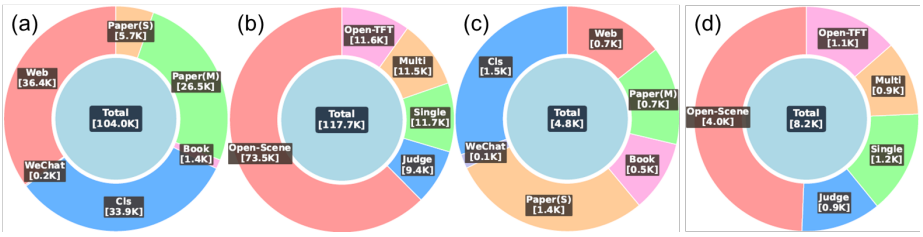

Fig. 2. Dataset statistics of IRIS-120K. (a)/(c) Distribution of image sources across Train/Test sets; (b)/(d) Distribution of generated VQA types (TFT vs. Scene-driven) in train and test, respectively.

3 Experiments

Experimental Setup. First, we fine-tune Qwen3-VL via Low-Rank Adaptation on four A6000 GPUs to get IRIS series models, more details can be found our repo. To comprehensively evaluate multi-modal capabilities, we benchmark IRIS against 16 contemporary VLMs, encompassing General-domain (GPT-5-mini [20], InternVL-3.5 series [5], Qwen3-VL series [2]) and Medical-domain models (HuatuoGPT-Vision [4], HuluMed [7], Lingshu [24], Med-Gemma [18]), which are evaluated in a zero-shot manner on the current test set. We employ Accuracy for closed-ended objective tasks (Judge, Single/Multi-choice) and standard generative metrics (BLEU-1 [16], ROUGE-1/L-F [11]) for open-ended interactions (Tree/Scene-VQA).

Main Results. Quantitative evaluations are summarized in Table 1. The IRIS demonstrates highly competitive performance across all question types. Our optimal model, **IRIS-4B**, achieves near-perfect objective reasoning accuracies (97.25 on Judge, 98.52 on Single-choice) and dominates complex generative tasks (e.g., 42.36 on Tree-VQA B-1). Specifically, across the five primary metrics (Accuracies and B-1 scores), IRIS-4B attains an outstanding overall average of **74.26**.

Crucially, IRIS exhibits profound parameter efficiency. IRIS-4B significantly outperforms baselines up to 8× larger, outperforming the strongest medical VLM (Lingshu-32B, overall avg. 55.00) by 19.26 absolute points, and the best generic VLM (Qwen3-VL-32B, overall avg. 50.61) by 23.65 points. Furthermore, even our minimalist IRIS-2B (avg. 69.57) consistently beats all 32B/34B competitors. Notably, the performance gap between IRIS-4B and its direct pretrained counterpart, Qwen3-VL-4B, effectively isolates the impact of domain-specific supervision within the same backbone family.

Ablation Studies. 1) Model Scaling: Evaluating the architecture across 2B, 4B, and 8B scales reveals that the performance peaks at 4B. Scaling further to 8B introduces a slight performance plateau, indicating that 4B is the optimal capacity sweet-spot for learning our specialized representations without overfitting, ideal for edge deployment. **2) Data Composition:** We ablated the training components of IRIS-4B. Training exclusively on structured TFT data (*IRIS-4B-TFT*) preserves strong logical deduction but degrades Scene-VQA gen-

Table 1. Performance comparison of General (Gen), Medical (Med), and Our models across all five question types. *Gen. Avg.* denotes the average score across the six open-ended generative metrics (**B**LEU and **R**OUGE). Best results are in **bold**.

Group	Model	Judge	Single	Multi	Tree-VQA			Scene-VQA			Gen.Avg.
					B-1	R-1-F	R-L-F	B-1	R-1-F	R-L-F	
Gen	GPT-5-mini	69.52	72.08	67.48	16.48	11.31	21.81	13.73	22.02	11.81	16.19
	Internvl3.5-14B	77.35	75.62	54.59	11.82	19.80	11.59	18.50	20.22	11.34	15.55
	Internvl3.5-8B	72.59	75.37	49.41	10.52	17.94	10.64	15.42	18.68	10.17	13.90
	Qwen3-vl-32B	77.57	79.98	62.64	13.65	20.79	13.06	19.22	20.36	11.84	16.49
	Qwen3-vl-8B	69.52	73.06	57.41	7.77	16.30	8.88	13.33	17.09	9.18	12.09
	Qwen3-vl-4B	62.43	70.68	51.98	6.19	14.24	7.83	10.86	16.39	8.63	10.69
	Qwen3-vl-2B	69.31	62.36	44.15	4.82	15.88	9.06	7.15	16.59	8.96	10.41
Med	HGPT-V-7B	32.70	25.62	8.45	21.34	22.65	15.02	19.58	18.53	11.59	18.12
	HGPT-V-34B	66.56	68.29	50.36	28.60	27.18	18.79	24.40	22.02	13.89	22.48
	HuluMed-7B	83.07	49.09	26.61	31.91	30.17	20.58	26.12	25.49	15.84	25.02
	HuluMed-14B	83.92	59.59	25.63	32.48	30.68	20.80	25.98	25.70	15.70	25.22
	HuluMed-32B	83.70	60.05	30.39	33.02	31.06	21.11	26.02	25.92	15.90	25.51
	Lingshu-7B	84.97	64.25	37.70	21.63	25.88	19.33	13.59	20.64	13.46	19.09
	Lingshu-32B	84.76	84.43	59.71	28.04	29.85	21.92	18.08	23.66	15.13	22.78
	MG-4B	66.88	60.38	36.22	16.13	22.68	16.52	17.76	19.94	12.27	17.55
	MG1.5-4B	71.01	58.24	31.14	11.52	17.19	9.98	11.72	16.36	8.26	12.51
Ours	IRIS-8B	97.04	98.19	88.55	42.27	**39.68**	**29.64**	44.19	**36.32**	**24.74**	**36.14**
	IRIS-4B	**97.25**	**98.52**	**88.96**	**42.36**	39.29	29.03	**44.23**	36.08	24.45	35.91
	IRIS-4B-TFT	96.51	98.19	88.26	41.86	39.18	28.95	40.46	28.86	18.17	32.91
	IRIS-4B-Scene	93.12	83.20	48.42	37.02	34.31	24.48	44.13	36.01	24.41	33.39
	IRIS-2B	96.19	97.03	83.98	40.40	38.06	27.77	30.28	26.70	16.86	30.01

eration. Conversely, relying solely on role-playing Scene data (*IRIS-4B-Scene*) precipitates a substantial degradation in objective precision (e.g., Multi-choice accuracy plummets from 88.96 to 48.42). This confirms the essential synergy of our framework: TFT builds the rigorous medical logic, while Scene-adaptation provides the clinical interaction interface.

Qualitative Analysis & Interpretability. Fig. 3 qualitatively validates the clinical reliability and transparency of IRIS. In complex closed-ended tasks (Fig. 3, top), IRIS-4B accurately diagnoses *Granular corneal dystrophy* and verifies stromal involvement, overcoming the diagnostic inconsistencies and hallucinations that plague much larger medical baselines (e.g., Lingshu-32B).

Crucially, for open-ended interactions (Fig. 3, bottom), IRIS-4B transcends "black-box" answering. Driven by our structured TFT training paradigm, it generates an explicit `<think>` block to articulate a rigorous clinical chain-of-thought—detailing visual observations and logical deductions—prior to formulating its final conclusion. Furthermore, IRIS-4B exhibits robust *visual grounding*. Generated clinical terms (e.g., "lesion", "adjacent to the lesion") are spatially anchored to the input image via attention maps (delineated by $> 25\%$ yellow and $> 50\%$ red importance contours). While functioning conceptually akin to Class Activation Mapping (CAM) rather than exact pixel-level segmentation,

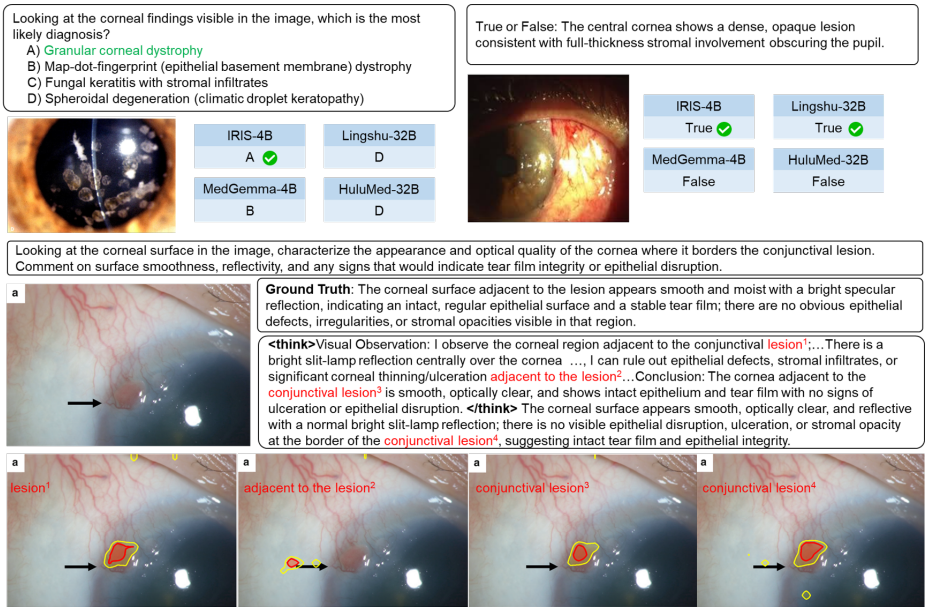

Fig. 3. Qualitative evaluation of IRIS-4B. **Top:** Accurate predictions on complex objective tasks, outperforming substantially larger medical baselines. **Bottom:** Demonstration of transparent reasoning and visual grounding. IRIS-4B articulates its logic via a `<think>` block and visually anchors generated clinical terms (red text) to localized pathological regions (attention contours: $> 25\%$ yellow, $> 50\%$ red).

this precise spatial-textual correlation confirms that IRIS-4B explicitly anchors its reasoning to authentic morphological cues rather than language priors, offering vital interpretability for reliable clinical triage.

4 Conclusion

In this paper, we introduced IRIS, a highly efficient and interpretable Vision-Language Model tailored for OSD screening. To overcome the severe domain gap in existing VLMs, we curated IRIS-120K, the largest external eye VQA dataset to date, and proposed a Clinically-driven Data Engine that combines a Topic Finding Tree with role-adaptive, scene-driven interactions. This design encourages the model to learn structured clinico-pathological reasoning rather than relying on superficial language priors. Extensive evaluations show that our optimal model, IRIS-4B, consistently outperforms general and medical baselines up to eight times its size across diverse multimodal tasks, supporting the key insight that high-quality, structurally curated domain data can outweigh sheer parameter scaling. Moreover, IRIS integrates explicit step-by-step reasoning chains (<think>) with attention-based visual grounding, improving interpretability, and aligning with clinical diagnostic workflows. Importantly, its effi-

ciency enables low-cost, fully offline edge deployment on patient devices, offering a practical path toward large-scale, privacy-preserving screening.

Despite these promising results, our study has several limitations. First, our evaluation is conducted in a closed-loop setting where the model is fine-tuned and evaluated on data splits generated by the same VLM/LLM-based data engine. Consequently, the reported results reflect benchmark-specific performance and distributional alignment rather than universal clinical superiority over all baselines. Second, while open-ended metrics (e.g., BLEU and ROUGE) are widely adopted to measure text alignment with references, they serve as weak proxies and are not complete proofs of diagnostic correctness or clinical safety. Moving forward, rigorous external clinical validation with expert multi-reader adjudication remains a necessary future work to confirm the real-world diagnostic reliability and practical utility of IRIS.

Acknowledgments. This work was supported in part by the Research Grants Council (RGC) of Hong Kong SAR (GRF14213125, GRF14201824, GRF14216222), the Innovation and Technology Fund (ITF) of Hong Kong SAR (ITS/252/23).

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

References

1. Baghbanzadeh, N., Fallahpour, A., Parhizkar, Y., Ogidi, F., Roy, S., Ashkezari, S., Khazaie, V.R., Colacci, M., Etemad, A., Afkanpour, A., et al.: Advancing medical representation learning through high-quality data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 24–33. Springer (2025)
2. Bai, S., Cai, Y., Chen, R., Chen, K., et.al, X.C.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Zhang, R., Cai, Z., Ji, K., et al.: Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. arXiv preprint arXiv:2406.19280 (2024)
5. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
6. Graham, A.D., Kothapalli, T., Wang, J., Ding, J., Tse, V., Asbell, P.A., Yu, S.X., Lin, M.C.: A machine learning approach to predicting dry eye-related signs, symptoms and diagnoses from meibography images. Heliyon **10**(17) (2024)
7. Jiang, S., Wang, Y., Song, S., Hu, T., Zhou, C., Pu, B., Zhang, Y., Yang, Z., Feng, Y., Zhou, J.T., et al.: Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. arXiv preprint arXiv:2510.08668 (2025)

8. Jin, K., Yu, T., Ying, G.s., Ge, Z., Li, K.Z., Zhou, Y., Shi, D., Wang, M., Goktas, P., Grzybowski, A.: A systematic review of vision and vision-language foundation models in ophthalmology. Advances in ophthalmology practice and research (2025)
9. Li, S., Lin, T., Lin, L., Zhang, W., Liu, J., Yang, X., Li, J., He, Y., Song, X., Xiao, J., et al.: Eyecaregpt: Boosting comprehensive ophthalmology understanding with tailored dataset, benchmark and model. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 3893–3902 (2025)
10. Li, Z., Wang, Z., Xiu, L., Zhang, P., Wang, W., Wang, Y., Chen, G., Yang, W., Chen, W.: Large language model-based multimodal system for detecting and grading ocular surface diseases from smartphone images. Frontiers in Cell and Developmental Biology **13**, 1600202 (2025)
11. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
12. Lozano, A., Sun, M.W., Burgess, J., Chen, L., Nirschl, J.J., Gu, J., Lopez, I., Aklilu, J., Rau, A., Katzer, A.W., et al.: Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific literature. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19724–19735 (2025)
13. Lu, T.C., Huang, C.H., Lin, I.C.: Artificial intelligence application in cornea and external diseases. Diagnostics **15**(24), 3199 (2025)
14. Maehara, H., Ueno, Y., Yamaguchi, T., Kitaguchi, Y., Miyazaki, D., Nejima, R., Inomata, T., Kato, N., Chikama, T.i., Ominato, J., et al.: Artificial intelligence support improves diagnosis accuracy in anterior segment eye diseases. Scientific Reports **15**(1), 5117 (2025)
15. Maehara, H., Ueno, Y., Yamaguchi, T., Kitaguchi, Y., Miyazaki, D., Nejima, R., Inomata, T., Kato, N., Chikama, T.i., Ominato, J., et al.: The importance of clinical experience in ai-assisted corneal diagnosis: verification using intentional ai misleading. Scientific Reports **15**(1), 1462 (2025)
16. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
17. Qiu, J., Wu, J., Wei, H., Shi, P., Zhang, M., Sun, Y., Li, L., Liu, H., Liu, H., Hou, S., et al.: Visionfm: a multi-modal multi-task vision foundation model for generalist ophthalmic artificial intelligence. arXiv preprint arXiv:2310.04992 (2023)
18. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
19. Shi, D., Zhang, W., Yang, J., Huang, S., Chen, X., Xu, P., Jin, K., Lin, S., Wei, J., Yusufu, M., et al.: A multimodal visual–language foundation model for computational ophthalmology. npj Digital Medicine **8**(1), 381 (2025)
20. Singh, A., Fry, A., Perelman, A., et al.: Openai gpt-5 system card (Dec 2025), `https://arxiv.org/abs/2601.03267`, accessed: 2026-02-21
21. Srinivasan, S., Ji, H., Chen, D.Z., Wong, W., Da Soh, Z., Goh, J.H.L., Pushpanathan, K., Wang, X., Ma, W., Wong, T.Y., et al.: Can off-the-shelf visual large language models detect and diagnose ocular diseases from retinal photographs? BMJ Open Ophthalmology **10**(1) (2025)
22. Sun, Y., Zhu, C., Zheng, S., Zhang, K., Sun, L., Shui, Z., Zhang, Y., Li, H., Yang, L.: Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5034–5042 (2024)

23. Wang, M.H., Pan, Y., Jiang, X., Lin, Z., Liu, H., Liu, Y., Cui, J., Tan, J., Gong, C., Hou, G., et al.: Leveraging artificial intelligence and clinical laboratory evidence to advance mobile health applications in ophthalmology: Taking the ocular surface disease as a case study. iLABMED **3**(1), 64–85 (2025)
24. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
25. Yeh, C.H., Wang, J., Graham, A.D., Liu, A.J., Tan, B., Chen, Y., Ma, Y., Lin, M.C.: Insight: A multi-modal diagnostic pipeline using llms for ocular surface disease diagnosis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 711–721. Springer (2024)
26. Yonehara, M., Nakagawa, Y., Ayatsuka, Y., Hara, Y., Shoji, J., Ebihara, N., Inomata, T., Huang, T., Nagino, K., Fukuda, K., et al.: Use of explainable ai on slit-lamp images of anterior surface of eyes to diagnose allergic conjunctival diseases. Allergology International **74**(1), 86–96 (2025)
27. Zauner, C.: Implementation and Benchmarking of Perceptual Image Hash Functions. Master's thesis, University of Applied Sciences Hagenberg, Hagenberg, Austria (Jul 2010), `https://www.phash.org/docs/pubs/thesis_zauner.pdf`, master's thesis. pHash library implementation
28. 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., et al.: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023)