

BACKGROUND

Retinal color fundus photography (CFP) is a widely used, non-invasive tool for diagnosing ocular and systemic diseases. With the rise of multimodal large language models (MLLMs) such as GPT-4 and LLaVA, there is increasing interest in applying joint visual-textual reasoning to medical imaging. While biomedical variants like LLaVA-Med have introduced domain-specific instruction tuning, current MLLMs still struggle to capture the structural complexity and subtle clinical features critical for retinal diagnosis.

OBJECTIVE

- ❖ Develop RetinalGPT, a retinal clinical assistant powered by large vision-language models.
- ❖ Integrate retinal images with structured clinical knowledge through an instruction fine-tuning pipeline.
- ❖ Enable quantitative vascular analysis, lesion localization, and disease diagnosis in an interpretable manner.

METHODS

We built a large-scale retinal dataset and designed a two-stage instruction tuning pipeline to adapt vision-language models for clinical retinal analysis.

- Data: Collected 38K fundus images (EyeQ, Messidor-1, APTOS, IDRiD, MICCAI MMAC, OIA-ODIR, RFMiD, and private), annotated with disease labels, lesion masks, vascular features (fractal dimension, vessel density, branching angles, curvature), and image quality (good/usable/reject) using AutoMorph and RBAD.
- Datasets: Constructed RCA (38K QA pairs) and RCT (38K multi-turn dialogues) via GPT-4.
- Training:
 - Stage 1 (Feature Alignment): 600K PMC-15M + 38K RCA, freeze vision encoder & LLM.
 - Stage 2 (Instruction Tuning): 38K RCT + 60K LLaVA-Med QA, freeze vision encoder, preserve general medical knowledge.

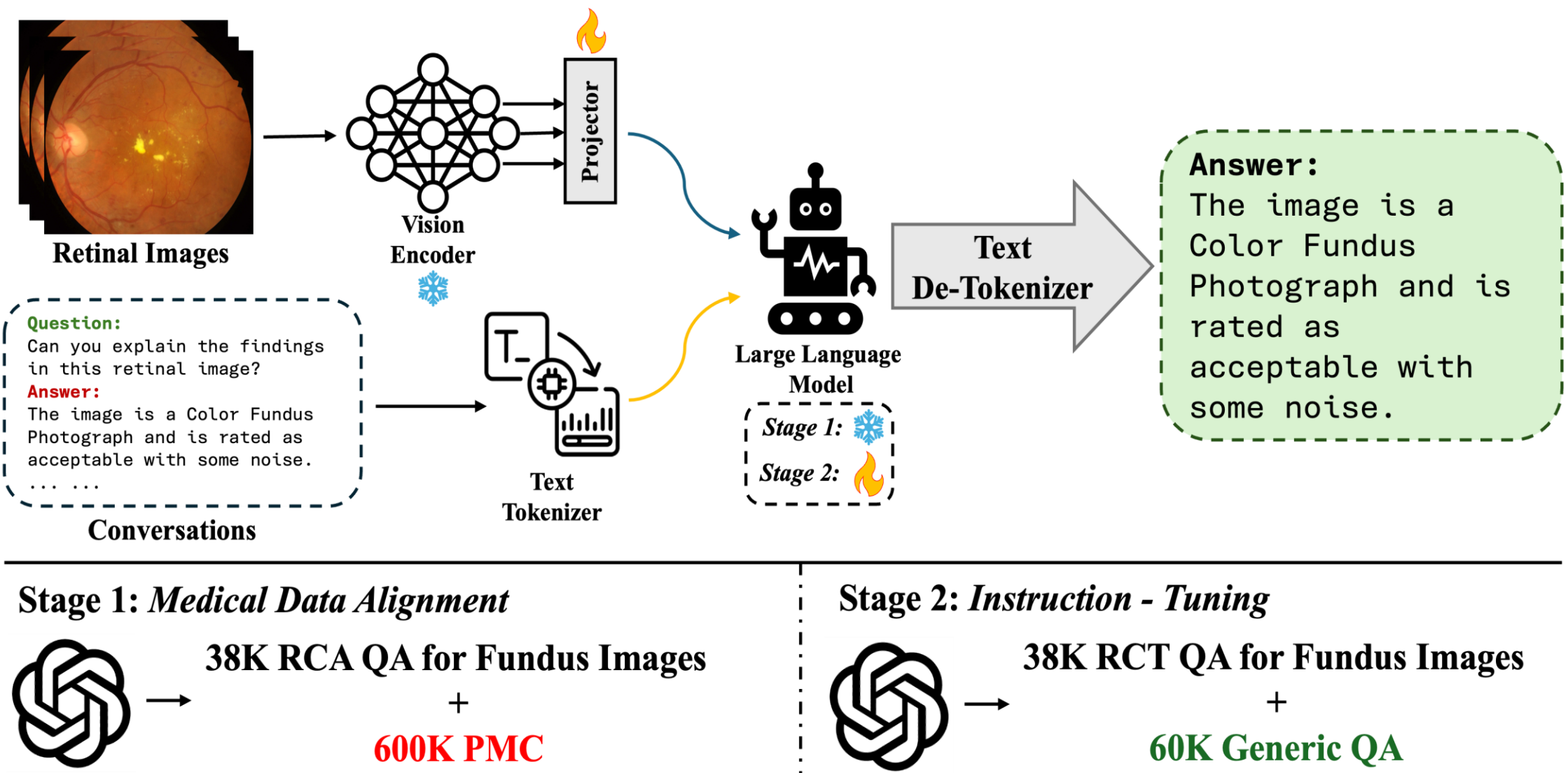


Fig.1 Overview of the model architecture and training strategy. Top: The network structure. Bottom: the two-stage training process.

RESULTS

- Superior accuracy: RetinalGPT outperforms GPT-4o, LLaVA, and LLaVA-Med across 8 retinal datasets, achieving up to 95%+ accuracy in DR grading and multi-disease classification.

Method	APTOS	EyePACS	IDRID	Messidor	MICCAI	OIA-ODIR	RFMiD	Private
Llama-3.2-11B-Vision	59.78	60.27	47.17	59.17	42.86	44.62	31.27	13.85
LLaVA	56.52	43.13	56.6	54.17	56.43	55.38	65.15	55.41
LLaVA-med	47.00	55.57	49.01	53.33	57.14	48.92	53.09	44.16
GPT-4o	54.35	53.03	52.83	58.33	45.71	52.69	58.63	31.17
Ours	95.10	73.70	84.90	80.83	87.14	66.13	88.27	99.57

Table 1. Performance comparison of different models on abnormality classification.

- Lesion localization (Generates bounding boxes) and Vascular analysis(fractal dimension, vessel density, and other morphological metrics).

User:

Can you tell me the location of lesions in the image?

Assistant:

Sure! Lesions are located in the following area: one bounding box is near the coordinates `[0.6538, 0.5112, 0.7198, 0.6312]`.

Lesion Localization

RED box is the Ground Truth
BLUE box is the prediction from Assistant

User:

Can you tell me about the vascular analysis results?

Assistant:

Sure! The fractal dimension of the retina vessels is **1.428**, the vessels density is **0.071**.

User:

Can you provide more details about the retinal veins?

Assistant:

The fractal dimension of the retinal veins is **1.236**, with a density of **0.032** and an average width of **14.078**.

User:

What about the analysis in Zone B?

Assistant:

In Zone B, the vessel density is **0.011**, the average vessel width is **15.901**, and the distance tortuosity is **1.721**.

Fractal Analysis of Vascular Structures

Vessels analysis(Ground Truth):

fractal demission of retina vessels is **1.445**, vessels density is **0.078**, average width of vessels is **14.171**, fractal dimension of retinal veins is **1.234**, density of retinal veins is **0.030**, average width of retinal veins is **14.274**, vessel density in Zone B is **0.012**, average vessel width in Zone B is **15.925**, distance tortuosity in Zone B is **1.383**

CONCLUSIONS

In this work, we introduce RetinalGPT, a multimodal large vision-language model designed for retinal image understanding. Leveraging a large-scale dataset, structured clinical annotations, and a two-stage instruction tuning strategy, RetinalGPT achieves accurate disease classification and lesion localization. Importantly, it can also provide vascular quantitative features, which support diagnosis in an interpretable and clinically meaningful way.

REFERENCES

- Li, Chunyuan, et al. "Llava-med: Training a large language-and-vision assistant for biomedicine in one day." Advances in Neural Information Processing Systems 36 (2023): 28541-28564.
- Liu, Haotian, et al. "Visual instruction tuning." Advances in neural information processing systems 36 (2023): 34892-34916.
- Zhou, Yukun, et al. "AutoMorph: automated retinal vascular morphology quantification via a deep learning pipeline." Translational vision science & technology 11.7 (2022): 12-12.
- Wang, Hao, et al. "Rbad: A dataset and benchmark for retinal vessels branching angle detection." 2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). IEEE, 2024.
- Zhang, Sheng, et al. "Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs." arXiv preprint arXiv:2303.00915 (2023).