



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

# Insight: A Multi-Modal Diagnostic Pipeline using LLMs for Ocular Surface Disease Diagnosis

Chun-Hsiao Yeh<sup>1,2</sup>, Jiayun Wang<sup>1,3</sup>, Andrew D. Graham<sup>1,2</sup>, Andrea J. Liu<sup>1</sup>, Bo Tan<sup>1</sup>, Yubei Chen<sup>4</sup>, Yi Ma<sup>2,5</sup>, and Meng C. Lin<sup>1,2</sup> (✉)

<sup>1</sup> Clinical Research Center, University of California, Berkeley, Berkeley, CA, USA

<sup>2</sup> University of California, Berkeley, Berkeley, CA, USA

{daniel\_yeh, mlin}@berkeley.edu

<sup>3</sup> California Institute of Technology, Pasadena, CA, USA

<sup>4</sup> University of California, Davis, Davis, CA, USA

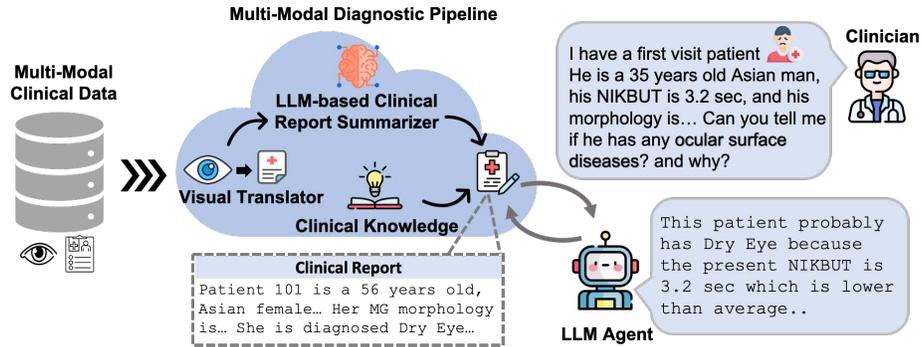
<sup>5</sup> University of Hong Kong, Hong Kong SAR, China

**Abstract.** Accurate diagnosis of ocular surface diseases is critical in optometry and ophthalmology, which hinge on integrating clinical data sources (e.g., meibography imaging and clinical metadata). Traditional human assessments lack precision in quantifying clinical observations, while current machine-based methods often treat diagnoses as multi-class classification problems, limiting the diagnoses to a predefined closed-set of curated answers without reasoning the clinical relevance of each variable to the diagnosis. To tackle these challenges, we introduce an innovative multi-modal diagnostic pipeline (MDPipe) by employing large language models (LLMs) for ocular surface disease diagnosis. We first employ a visual translator to interpret meibography images by converting them into quantifiable morphology data, facilitating their integration with clinical metadata and enabling the communication of nuanced medical insight to LLMs. To further advance this communication, we introduce a LLM-based summarizer to contextualize the insight from the combined morphology and clinical metadata, and generate clinical report summaries. Finally, we refine the LLMs' reasoning ability with domain-specific insight from real-life clinician diagnoses. Our evaluation across diverse ocular surface disease diagnosis benchmarks demonstrates that MDPipe outperforms existing standards, including GPT-4, and provides clinically sound rationales for diagnoses. The project is available at <https://danielchye.github.io/MDPipe/>.

**Keywords:** Multimodality · Large Language Models · Ocular Surface Disease Diagnosis

## 1 Introduction

Ocular surface diseases (OSD) are various disorders impacting the anterior segment of the eye, with Dry Eye (DE) being the most prevalent. DE significantly impacts ocular surface health, vision, and quality of life, and is a predominant reason for eye care visits globally [4,20]. Evaporative DE, the most common



**Fig. 1.** Multi-modal diagnostic pipeline using LLMs for OSD diagnosis. The proposed pipeline utilizes 1) a visual translator to transform meibography images into quantifiable MG morphology, 2) an LLM-based summarizer to craft clinical reports, and 3) the integration of clinical knowledge to augment LLM’s capability in diagnosing OSD.

subtype, is primarily attributed to Meibomian Gland Dysfunction (MGD) [14] which is characterized by the glands’ failure to secrete a sufficiently thick and well-organized lipid layer [3]. This results in increased tear evaporation, tear film thinning and destabilization, leading to tear hyperosmolarity and premature tear breakup, culminating in the discomfort associated with DE [24]. Traditional diagnosis involves clinical assessments such as measurement of fluorescein tear breakup time (FTBUT), noninvasive keratograph breakup time (NIKBUT), grading of MG expressate, meibography imaging, as well as administration of symptom instruments such as the Ocular Surface Disease Index (OSDI) [19]. Yet, these approaches are time-consuming [27] and have poor reproducibility [15], particularly in quantifying meibomian gland (MG) atrophy [6,1] where estimations are subject to human bias and lack the accuracy for a definitive diagnosis.

Recently, efforts to employ machine learning (ML) in the diagnosis of OSD have seen notable advancements. Initial efforts have concentrated on quantifying MG morphology through imaging techniques [21,30] and segmentation models for MG feature extraction [28,18,27]. As research evolves, there is a growing emphasis on combining visual assessments of the MGs with clinical metadata [11]. However, incorporating a diverse set of clinical variables into predictive classification models results in closed-set predictions based on predefined sets of answers, and treats clinical variables merely as data points without semantic relationships or diagnostic rationale. Thus, understanding the clinical implications of the metadata and how these variables relate to conditions such as DE remains an ongoing challenge that requires further exploration.

The emergence of Multimodal Large Language Models (MLLMs) [16,34,13] brings new light to this problem. MLLMs leverage the powerful abilities of LLMs [17,26] while integrating visual data. MLLMs, when utilized for diagnostic purposes [9,12,33], are capable of producing hundreds to thousands of free-form answers with clinical reasoning, instead of being limited to closed-set predictions.

However, several critical analyses of MLLMs reveal persistent challenges in accurately representing visual data [25,32]. This issue becomes particularly pronounced when dealing with rare domains or complex imagery, such as meibography (see Fig. 2).

Addressing this gap, we pose two questions: 1) Can a model process meibography images with the same level of attention and detail as a human clinician? and 2) Can the model make a precise and accurate diagnosis by focusing on specific MG morphological features and quantifying those observations along with clinical metadata? To answer these questions, we propose a novel multi-modal diagnostic pipeline (MDPipe) that integrates patient clinical metadata and visual data for OSD diagnosis through the use of LLMs (see Fig. 1). Specifically, we contribute in three major aspects:

- We propose a visual translator to convert meibography images into quantifiable MG morphology data through an instance segmentation network. This transformation allows us to bridge the visual data with clinical metadata, effectively marrying detailed morphological insights with clinical context.
- To advance this integration, we introduce an LLM-based summarizer to generate clinical report summaries to contextualize the insights from both the non-narrative clinical metadata and the MG morphology data.
- We further collect clinical knowledge using real-life clinician diagnoses to refine the LLMs with nuanced, domain-specific knowledge.

## 2 Method

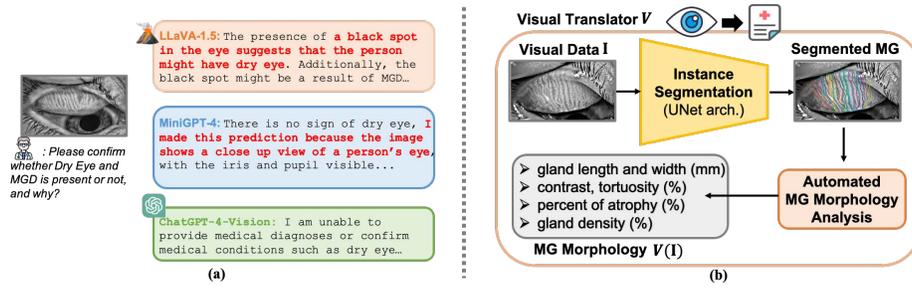
### 2.1 Problem Formulation

Our objective is to develop a multi-modal pipeline integrating patient clinical metadata  $\mathbf{M} = (\mathbf{C}, \mathbf{D})$  where  $\mathbf{C}$  represents clinical measurements and  $\mathbf{D}$  denotes the clinician’s diagnosis, with meibography images  $\mathbf{I}$ . This integration aims to use LLMs to enhance ocular disease diagnostics by fine-tuning an LLM  $p_\theta$  to diagnose patients using summarized clinical data.

We first use a report summarizer  $\mathcal{S}$  implemented using GPT-4 [16] to convert the metadata  $\mathbf{M}$  and image data  $\mathbf{I}$  into question-answer pairs  $(\mathbf{Q}, \mathbf{A}) = \mathcal{S}(\mathbf{M}, \mathcal{V}(\mathbf{I}))$ , where  $\mathcal{V}$  is a visual translator (in practice, implemented using a pre-trained segmentation model) that extracts MG morphology from an image  $\mathbf{I}$ . The ultimate goal is to refine the LLM parameters  $\theta^*$  through the maximization of the conditional log-likelihood, formalized as:  $\theta^* = \arg \max_\theta \sum_{i=1}^N \log p_\theta(\mathbf{A}_i | \mathbf{Q}_i)$ , where  $N$  is the total number of cases. In practice, given new clinical data  $\mathbf{C}^*$  and image  $\mathbf{I}^*$ , we can generate a new query  $\mathbf{Q}^* = \mathcal{S}(\mathbf{C}^*, \mathcal{V}(\mathbf{I}^*))$ , which can then be fed into the LLMs to predict the corresponding diagnosis  $\mathbf{A}^*$ .

### 2.2 Visual Translator

MLLMs have known limitations when processing visual data and generating corresponding representations. The intricate details captured in meibography



**Fig. 2.** (a) Illustration of the limitations of current MLLMs in processing visual data, including: 1) producing vague interpretations; 2) not delivering clinical significance for ocular surface diseases, such as labeling "a black spot in the eye". (b) Our visual translator  $\mathcal{V}$  is designed to interpret visual data  $I$  by converting them into quantifiable MG morphology data.

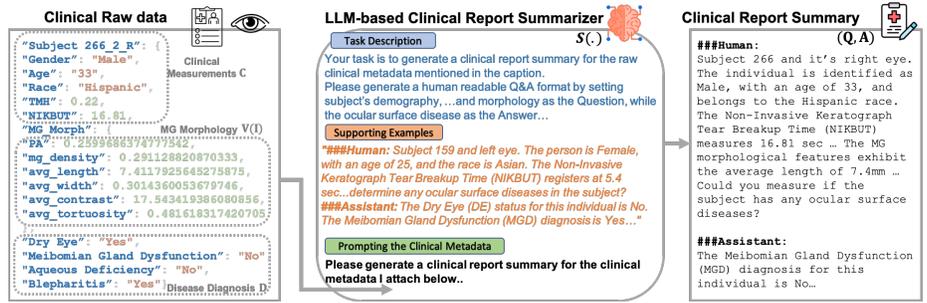
images present a particular challenge (see Fig. 2-(a)) due to their complexity and the nuanced information they hold. To address this, we set out to construct a model to process meibography images with the same level of attention and detail as a human clinician and to quantify the observations. Building on prior works [5,27], we introduced a visual translator  $\mathcal{V}$ , designed to convert meibography images  $I$  into quantifiable MG morphology data. This transformation served as a bridge, linking the raw visual data with clinical metadata and thus weaving together rich morphological detail with pertinent clinical context.

The visual translator operated through a two-step process. (see Fig. 2-(b)) Firstly, meibography images were passed to an instance segmentation network [5] that automatically delineated individual gland regions. Secondly, a detailed quantification of morphological features at the MG level was performed. By automating these steps, the model enhanced the observational capabilities of clinicians, providing a quantitative assessment of the glandular characteristics that were emerging as important factors in diagnosing OSD.

With the visual translator, we were able to precisely measure morphological features such as percent atrophy, gland density, average gland local contrast, and dimensions of gland length, width, and tortuosity [31]. These measures were quantified, allowing for an accurate depiction of the MG morphology, beyond what manual estimation could achieve.

### 2.3 LLM-based Clinical Report Summarizer

As illustrated in Fig. 3 (left), the clinical metadata and MG morphology, derived via the visual translator, presented in a discrete and non-narrative form. Our goal was to input this fragmented data into GPT-4 [16] to synthesize a cohesive clinical report summary for each case. The prompting structure provided to GPT-4, depicted in Fig. 3 (middle), had three main components: 1) a task description, 2) supporting examples, and 3) prompting the clinical metadata. Each component



**Fig. 3.** We employed an LLM-based summarizer to generate Q&A clinical reports to contextualize insights from both the non-narrative clinical metadata and MG morphology to enhance LLMs’ learning capability.

was designed to guide GPT-4 toward generating an informative and clinically relevant summary, which we delineated as follows:

**Task Description.** We first directed the model to perform the task of generating a clinical report summary from clinical raw data. The instruction specified the Q&A format, with the subject’s demographics, clinical, and morphological features data forming the questions, and the suspected OSD as the answer. This concise directive ensured that the output was clinically informative and formatted.

**Supporting Examples.** Drawing from [10], we enhanced the model’s understanding by supplementing the task description with manually curated examples. These examples served to illustrate the desired approach for crafting clinical report summaries and were strategically employed to resolve ambiguities, guiding GPT-4 toward producing outputs that aligned with clinical reporting standards.

**Prompting the Clinical Metadata.** Upon presenting the supporting examples, we proceeded to prompt the model with a request to generate a clinical report summary, attaching the raw clinical metadata subsequently. This approach enabled the model to synthesize a clinical report summary that adhered closely to our predefined instructions. Template can be found in the Appendix.

## 2.4 Refinement of LLMs with Clinical Knowledge

**Fundamental Clinical Knowledge.** To refine an LLM for ocular disease diagnosis, we enhanced it with targeted knowledge using DE-specific clinical terms and definitions extracted from 15 published DE clinical trials conducted between 2015 and 2023 [23,22]. This information was formatted into Q&A pairs with GPT-4, emphasizing common trial criteria like specific FTBUT ranges and OSDI score thresholds. Our fine-tuning process re-focused the LLM on key

medical data from DE clinical trials, potentially already within its training set, to deepen its clinical understanding.

**Real-Life Clinician Diagnoses.** For the second refinement phase, we integrated authentic clinical diagnostic cases to ground the LLMs in empirical medical expertise. A dataset comprising 20 real-world cases was compiled by human clinicians, encompassing both the clinicians’ diagnosis of various OSD (e.g., MGD, blepharitis), and the corresponding rationale underpinning these clinical judgments. This dataset was effectively transformed into a series of Q&A pairs, effectively inculcating the LLMs with the nuanced reasoning and decision-making processes typical of seasoned clinicians.

### 3 Experimental Setup

**Datasets.** We utilized two multimodal datasets: CRC [28,27] and DREAM [2,8], which included meibography images and clinical metadata. CRC is a curated dataset that spans a spectrum from normal to moderate DE conditions. Conversely, the DREAM dataset primarily contains data from patients with moderate to severe DE conditions. We merged these two datasets and, via our processing pipeline, converted them into a QA-based clinical report format conducive to diagnostic tasks. The resultant composite dataset contained a total of 3513 entries. Train/test split is 90%/10%. Training set has 1903 metadata-only and 1257 image+metadata instances; Test set has 198 metadata-only and 155 image+metadata instances. There are a total of 878 subjects.

**Metric.** For quantitative analysis, we followed established benchmarks [29], utilizing accuracy and F1 score. We also incorporated sensitivity (SN) and specificity (SP) metrics. On the qualitative front, we engaged human clinicians in a user study to grade the LLM’s diagnostic outputs across multiple dimensions.

**Implementation Details.** We chose the 7B and 13B versions of pre-trained LLaMA-2 [26]. We obtained the model checkpoints from the official Huggingface (NousResearch/Llama-2-7b-chat-hf, NousResearch/Llama-2-13b-chat-hf). All models were able to train on four NVIDIA Geforce RTX 3090 GPUs (average training time: ~8 hrs). We used the AdamW optimizer with 0.03 warm-up ratio and a learning rate of  $2e-4$ . More details can be found in the Appendix.

## 4 Results

**Comparison of General and Medical LLMs.** In Table 1, we evaluate the performance of various LLMs in diagnosing ocular diseases, specifically Dry Eye (DE), Meibomian Gland Dysfunction (MGD), and Blepharitis. Notably, GPT-4 demonstrates consistent performance across all three ocular diseases.

**Table 1.** Comparison between general and medical domain-tuned LLMs for diagnosing ocular diseases: Dry Eye (DE), Meibomian Gland Dysfunction (MGD), and Blepharitis. Evaluation criteria include accuracy, sensitivity (SN), specificity (SP), and F1 score. Our MDPipe outperforms existing benchmarks across three ocular diseases.

Method / Disease	DE				MGD				Blepharitis			
	Acc.	SN	SP	F1	Acc.	SN	SP	F1	Acc.	SN	SP	F1
<i>General LLMs without fine-tuning</i>												
Llama [26]	49.8	93.2	14.7	60.5	40.6	88.7	17.1	55.9	44.7	28.5	55.3	30.8
GPT-3.5 [17]	57.7	86.7	32.7	64.9	48.6	95.5	25.6	60.6	49.2	31.3	61.9	33.8
Llama2-7B [26]	63.9	88.2	38.6	66.6	52.7	83.2	23.3	62.3	47.4	31.8	59.3	34.4
GPT-4 [16]	70.7	77.1	66.3	67.7	65.2	65.7	76.8	65.5	58.2	39.3	72.9	48.8
<i>LLMs fine-tuned on medical domain data</i>												
Med-Alpaca [7]	62.5	87.3	33.5	70.3	53.4	84.7	28.2	61.9	54.9	53.8	55.8	49.7
PMC-LLaMA [29]	73.3	73.1	77.7	75.8	63.6	70.7	61.5	64.7	60.5	50.3	74.4	56.8
MDPipe-7B (ours)	86.9	89.3	84.3	87.8	<b>76.1</b>	<b>67.2</b>	<b>81.7</b>	<b>69.2</b>	71.2	56.3	79.7	63.8
MDPipe-13B (ours)	<b>89.5</b>	<b>88.2</b>	<b>91.0</b>	<b>89.9</b>	74.4	61.4	82.9	65.7	<b>73.1</b>	<b>58.7</b>	<b>80.1</b>	<b>65.1</b>

This proficiency is likely attributed to its extensive pre-training across a broad dataset, encompassing a range of medical knowledge, which may contribute to its enhanced ability to generalize across different domains.

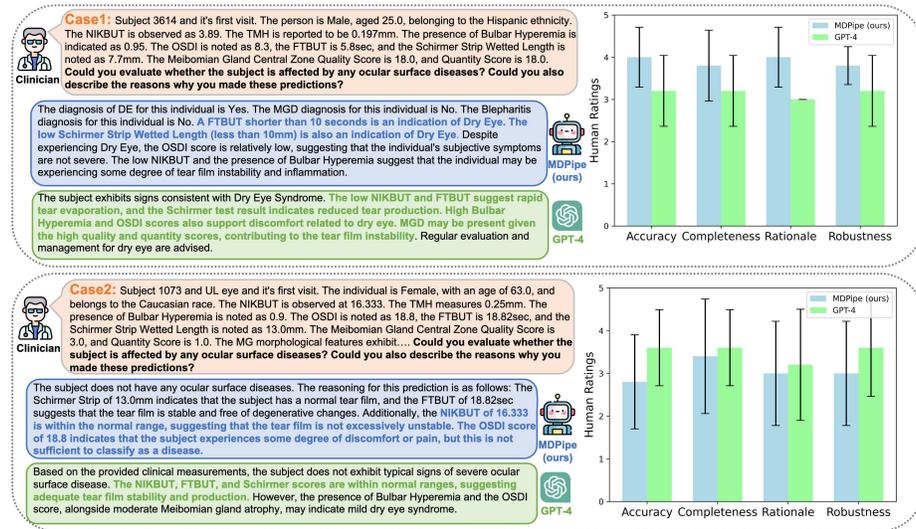
When considering models fine-tuned on medical domain data, PMC-LLaMA showed notable improvements over general LLMs, achieving an accuracy of 73.3% for DE, demonstrating the advantage of domain-specific fine-tuning. However, our MDPipe models significantly outperform all other models, achieving an accuracy of 86.9% for DE, 81.7% for MGD, and 79.7% for Blepharitis.

**Ablation Studies.** In Table 2, we investigated the impact of various training variables within our MDPipe on the diagnostic accuracy of the pre-trained LLaMA2-7B model. The variables considered were clinical metadata, MG morphology data, MG expression (quality and quantity scores), and real clinician diagnoses. We found the inclusion of MG morphology notably improved performance, especially in MGD diagnosis, increasing accuracy from 65.5% to 74.4%. This suggests the significant role of morphological features in detecting MGD.

**User (Clinician) Preference Study.** A user (clinician) preference study was performed to evaluate the performance of our MDPipe and GPT-4 in diagnosing OSD. A series of three random cases were presented with the appropriate diagnoses and rationale. Specifically, 5 clinicians were masked as to which model produced each output, and then asked to read and rate the two models' output on a scale from 1 (poor) to 5 (best) regarding 1) clinical accuracy, 2) diagnostic

**Table 2.** The impact of various training variables within our MDPipe on ocular disease diagnosis. Notably, it is observed that MG morphology is essential in MGD diagnosis.

Pretrain	+ Training Variables in MDPipe				Diagnosis Acc. (%)		
	Metadata	Morphology	MG-Express.	Real Diag.	DE	MGD	Bleph.
	✓	✗	✗	✗	83.5	65.5	69.4
LLaMA2	✓	✓	✗	✗	84.1	74.4	68.8
	✓	✓	✓	✗	85.8	75.6	70.1
	✓	✓	✓	✓	<b>86.9</b>	<b>76.1</b>	<b>71.2</b>

**Fig. 4.** Comparative evaluation and clinician study between MDPipe and GPT-4.

completeness, 3) diagnostic rationale, and 4) the model's robustness to handle ambiguous or incomplete patient data.

Figure 4 depicts the comparative study between MDPipe and GPT-4 using two clinical cases. In case 1, our MDPipe surpasses GPT-4, as evidenced by higher scores in accuracy and rationale. MDPipe's diagnostic responses apply DE criteria (e.g., FTBUT < 10 sec and Schirmer's test measurements < 10mm), which suggests a successful incorporation of DE clinical trials into the LLMs as detailed in Section 2.4. Conversely, GPT-4 seems to lack a nuanced interpretation of clinical measurements. For instance, one clinician remarks *"The OSDI is <13 and normal, I would not consider 0.95 bulbar hyperemia to be signs of dry eye"*, implying a potential gap in GPT-4's understanding of the subtleties within clinical scale values and their association with disease states. In case 2, MDPipe exhibits shortcomings in completeness, as noted by a clinician's feedback, *"The*

*diagnosis does not take into account MG expression and morphological data...*" pointing out the omission of some variables, indicating a potential oversight of comprehensive variable consideration in the diagnostic process.

## 5 Conclusion

MDPipe makes a significant novel contribution to ocular disease diagnosis by harnessing LLMs to address the limitations of current diagnostic methods. We integrate a visual translator that quantifies the visual data and clinician expertise for robust reasoning. As evidenced by benchmarking and the user study, MDPipe demonstrates a significant improvement in accurate OSD diagnosis with clinically relevant rationale. For future work, expanding the dataset of real-life clinician diagnoses and clinically verified variable scales promises to further refine the robustness and depth of LLM-generated rationales.

**Acknowledgments.** This work is supported by NIH (R21EY033881); UC Berkeley, Clinical Research Center Unrestricted Fund; and the Roberta J. Smith Research Fund.

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

## References

1. Arita, R., Itoh, K., Maeda, S., Maeda, K., Furuta, A., Fukuoka, S., Tomidokoro, A., Amano, S.: Proposed diagnostic criteria for obstructive meibomian gland dysfunction. *Ophthalmology* **116**(11), 2058–2063 (2009) [2](#)
2. Asbell, P.A., Maguire, M.G., Peskin, E., Bunya, V.Y., Kuklinski, E.J., et al.: Dry eye assessment and management (dream©) study: study design and baseline characteristics. *Contemporary Clinical Trials* **71**, 70–79 (2018) [6](#)
3. Bron, A., Tiffany, J., Gouveia, S., Yokoi, N., Voon, L.: Functional aspects of the tear film lipid layer. *Experimental eye research* **78**(3), 347–360 (2004) [2](#)
4. Craig, J.P., Nichols, K.K., Akpek, E.K., Caffery, B., Dua, H.S., Joo, C.K., Liu, Z., Nelson, J.D., Nichols, J.J., Tsubota, K., et al.: Tfos deus ii definition and classification report. *The ocular surface* **15**(3), 276–283 (2017) [1](#)
5. De Brabandere, B., Neven, D., Van Gool, L.: Semantic instance segmentation with a discriminative loss function. *arXiv preprint arXiv:1708.02551* (2017) [4](#)
6. Finis, D., Ackermann, P., Pischel, N., König, C., Hayajneh, J., Borrelli, M., Schrader, S., Geerling, G.: Evaluation of meibomian gland dysfunction and local distribution of meibomian gland atrophy by non-contact infrared meibography. *Current eye research* **40**(10), 982–989 (2015) [2](#)
7. Han, T., Adams, L.C., Papaioannou, J.M., Grundmann, P., Oberhauser, T., Löser, A., Truhn, D., Bressemer, K.K.: Medalpaca—an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247* (2023) [7](#)
8. Hussain, M., Shtein, R.M., Pistilli, M., Maguire, M.G., Oydanich, M., Asbell, P.A., Group, D.S.R., et al.: The dry eye assessment and management (dream) extension study—a randomized clinical trial of withdrawal of supplementation with omega-3 fatty acid in patients with dry eye disease. *The ocular surface* **18**(1), 47–55 (2020) [6](#)

9. 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. *Advances in Neural Information Processing Systems* **36** (2024) [2](#)
10. Lian, L., Li, B., Yala, A., Darrell, T.: Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *arXiv preprint arXiv:2305.13655* (2023) [5](#)
11. Lin, M.C., Graham, A.D., Kothapalli, T., Wang, J., Ding, J., Tse, V., Stella, X.Y.: Lifestyle and behaviors: predicting clinical signs and symptoms with machine learning. *Investigative Ophthalmology & Visual Science* **64**(8), 2880–2880 (2023) [2](#)
12. Liu, C., Cheng, S., Chen, C., Qiao, M., Zhang, W., Shah, A., Bai, W., Arcucci, R.: M-flag: Medical vision-language pre-training with frozen language models and latent space geometry optimization. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 637–647. Springer (2023) [2](#)
13. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. *arXiv preprint arXiv:2304.08485* (2023) [2](#)
14. Nichols, K.K., Foulks, G.N., Bron, A.J., Glasgow, B.J., Dogru, M., Tsubota, K., Lemp, M.A., Sullivan, D.A.: The international workshop on meibomian gland dysfunction: executive summary. *Investigative ophthalmology & visual science* **52**(4), 1922–1929 (2011) [2](#)
15. Nichols, K.K., Mitchell, G.L., Zadnik, K.: The repeatability of clinical measurements of dry eye. *Cornea* **23**(3), 272–285 (2004) [2](#)
16. OpenAI: Gpt-4 technical report (2023) [2](#), [3](#), [4](#), [7](#)
17. OpenAI, R.: Gpt-4 technical report. arxiv 2303.08774. View in Article [2](#), [13](#) (2023) [2](#), [7](#)
18. Prabhu, S.M., Chakiat, A., Shashank, S., Vunnava, K.P., Shetty, R.: Deep learning segmentation and quantification of meibomian glands. *Biomedical signal processing and control* **57**, 101776 (2020) [2](#)
19. Pult, H., Purslow, C., Murphy, P.J.: The relationship between clinical signs and dry eye symptoms. *Eye* **25**(4), 502–510 (2011) [2](#)
20. Rouen, P.A., White, M.L.: Dry eye disease: prevalence, assessment, and management. *Home healthcare now* **36**(2), 74–83 (2018) [1](#)
21. Saha, R.K., Chowdhury, A.M., Na, K.S., Hwang, G.D., Eom, Y., Kim, J., Jeon, H.G., Hwang, H.S., Chung, E.: Automated quantification of meibomian gland dropout in infrared meibography using deep learning. *The Ocular Surface* **26**, 283–294 (2022) [2](#)
22. Schmidl, D., Bata, A.M., Szegedi, S., Aranha Dos Santos, V., Stegmann, H., Fondi, K., Krösser, S., Werkmeister, R.M., Schmetterer, L., Garhöfer, G.: Influence of perfluorohexyloctane eye drops on tear film thickness in patients with mild to moderate dry eye disease: a randomized controlled clinical trial. *Journal of Ocular Pharmacology and Therapeutics* **36**(3), 154–161 (2020) [5](#)
23. Tauber, J., Wirta, D.L., Sall, K., Majmudar, P.A., Willen, D., Krösser, S., Group, S.S., et al.: A randomized clinical study (seecase) to assess efficacy, safety, and tolerability of nov03 for treatment of dry eye disease. *Cornea* **40**(9), 1132–1140 (2021) [5](#)
24. Teo, C.H.Y., Ong, H.S., Liu, Y.C., Tong, L.: Meibomian gland dysfunction is the primary determinant of dry eye symptoms: Analysis of 2346 patients. *The Ocular Surface* **18**(4), 604–612 (2020) [2](#)
25. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209* (2024) [3](#)

26. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023) [2](#), [6](#), [7](#)
27. Wang, J., Li, S., Yeh, T.N., Chakraborty, R., Graham, A.D., Stella, X.Y., Lin, M.C.: Quantifying meibomian gland morphology using artificial intelligence. *Optometry and Vision Science* **98**(9), 1094–1103 (2021) [2](#), [4](#), [6](#)
28. Wang, J., Yeh, T.N., Chakraborty, R., Stella, X.Y., Lin, M.C.: A deep learning approach for meibomian gland atrophy evaluation in meibography images. *Translational vision science & technology* **8**(6), 37–37 (2019) [2](#), [6](#)
29. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Pmc-llama: Further finetuning llama on medical papers. arXiv preprint arXiv:2304.14454 (2023) [6](#), [7](#)
30. Yeh, C.H., Stella, X.Y., Lin, M.C.: Meibography phenotyping and classification from unsupervised discriminative feature learning. *Translational vision science & technology* **10**(2), 4–4 (2021) [2](#)
31. Yeh, T.N., Lin, M.C.: Repeatability of meibomian gland contrast, a potential indicator of meibomian gland function. *Cornea* **38**(2), 256 (2019) [4](#)
32. Zhai, Y., Tong, S., Li, X., Cai, M., Qu, Q., Lee, Y.J., Ma, Y.: Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In: *Conference on Parsimony and Learning*. pp. 202–227. PMLR (2024) [3](#)
33. Zhang, K., Yu, J., Yan, Z., Liu, Y., Adhikarla, E., Fu, S., Chen, X., Chen, C., Zhou, Y., Li, X., et al.: Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. arXiv preprint arXiv:2305.17100 (2023) [2](#)
34. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023) [2](#)