MoEA-Net: Modality-Incremental Expert Aggregation Network for Retinal Prognostic Prediction

Hua Wang[1], Xiaodan Zhang[1], Yanzhao Shi[2*], Chengxin Zheng[1], Wanyu Zhang[1], Zhen Wang[1]
Jianing Wang[3], Xiaobing Yu[3]

[1]College of Computer Science, Beijing University of Technology, Beijing, China
[2]Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China
[3]Department of Ophthalmology, Beijing Hospital, National Center of Gerontology, Institute of Geriatric Medicine, Chinese Academy of Medical Sciences, Beijing, China
wanghua@emails.bjut.edu.cn, zhangxiaodan@bjut.edu.cn,yanzhaoshi@connect.hku.hk

Abstract

Automated analysis of temporal changes in multimodal retinal images is critical for the prognostic assessment of ophthalmic diseases. Unlike traditional single-timepoint diagnosis, tracking longitudinal changes across multiple imaging modalities introduces significant data bias challenges: (1) Imbalanced modality samples compromise the integration of knowledge within minority modalities; (2) Heterogeneous visual patterns across modalities undermine the perception of disease-relevant biomarkers. To tackle these issues, we propose a Modality-Incremental Expert Aggregation Network (MoEA-Net), which unifies the inter-modal integration and intra-modal perception for enhanced retinal prognostic prediction. Specifically, we employ the large language model (LLM) with incremental LoRA layers for specific modalities to effectively integrate knowledge from imbalanced data. Besides, we introduce a Spatiotemporal-aware Expert (SAE) module to better perceive both the anatomical structures and longitudinal changes within modalities. By progressively combining the SAE module with incremental LoRA, MoEA-Net supports continual knowledge accumulation and improves accurate reasoning. Experimental results show that MoEA-Net achieves state-of-the-art performance on *subretinal fluid change* and *visual recovery* classification tasks, validating its effectiveness.

Code — https://github.com/TommyK1ng/MoEA-Net

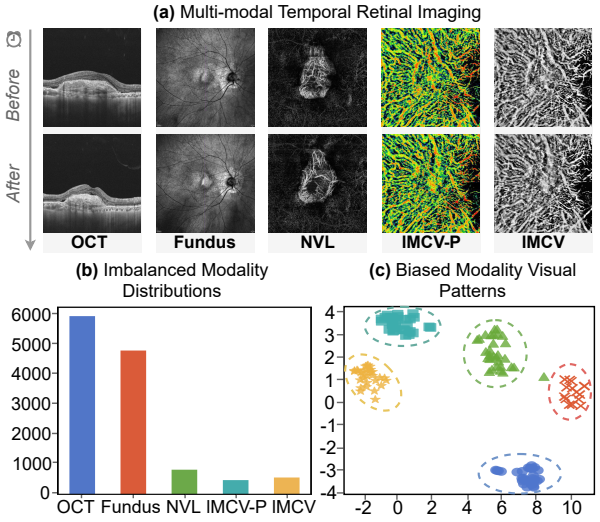

Figure 1: (a) Examples of multi-modal retinal images before and after treatment. (b) Illustration of imbalanced modality data in the dataset. (c) t-SNE (Van der Maaten and Hinton 2008) visualization of visual features from different modalities, highlighting substantial inter-modality variance.

1 Introduction

Retinal prognosis is vital for tracking disease progression in conditions such as age-related macular degeneration (AMD) and diabetic macular edema (DME). Clinicians typically assess prognosis by tracking temporal changes in retinal fluid and visual function, relying on various imaging modalities such as Optical Coherence Tomography (OCT) and fundus photography to interpret complex pathological patterns (Watanabe et al. 2022; Hao, Liu, and Yu 2022).

However, this manual process is time-consuming and prone to observer variability and error, especially when tracking subtle changes indicative of disease progression. Automated analysis of retinal temporal changes holds considerable promise for improving diagnostic accuracy and streamlining clinical workflows. Recent advances in deep neural networks and multi-modal large language models (MLLMs) have revolutionized optical image analysis. Early studies primarily focus on single-modality inputs (e.g., OCT or fundus images) to infer ophthalmic conditions (Liefers et al. 2017; Li et al. 2021b; He et al. 2023). However, this approach neglects complementary details present in other modalities, resulting in unreliable outcomes. To mitigate this issue, recent works explore modality fusion techniques to integrate information across modalities with varying resolutions and textures. Representative approaches include dual-stream fusion (Li et al. 2021a; Shafiq et al. 2024), knowledge distillation (Wang et al. 2023, 2025), and MLLM-based reasoning (Zhao et al. 2023). Such approaches enrich the model

*Corresponding Author

with complementary features, leading to more robust visual representations and enhanced classification performance.

Despite recent advances, existing methods for modeling temporal retinal changes across multiple modalities suffer from severe modality data bias. (1) **Imbalanced modality distributions**: As shown in Figure 1(b), sample counts vary widely across modalities, causing models to overfit dominant sources (e.g., OCT) while underutilizing rare ones, e.g., Non-vascular Layer Blood Flow Imaging (NVL). Although teacher-student frameworks (Wang et al. 2023, 2025) and LLM-based multi-modal prompting (Zhao et al. 2023) offer partial solutions, they lack flexibility to dynamically integrate newly available modalities, thereby limiting their practical utility in clinical settings. (2) **Heterogeneous visual patterns**: As visualized in Figure 1(a,c), different modalities exhibit distinct visual textures and representations. This complicates the model's ability to track disease-relevant biomarkers across time, leading to an insufficient perception of subtle retinal changes, thereby degrading performance. Thus, how to effectively integrate and perceive biased retinal modality data is an open question for prognostic prediction.

In this paper, we present the Modality-Incremental Expert Aggregation Network (MoEA-Net), a unified framework designed to enhance retinal prognostic prediction by effective inter-modal integration and intra-modal perception. It provides flexibility for clinical applications where retinal modalities are incomplete and uneven in number. The core idea of MoEA-Net is to harness the generalization capacity of LLMs for incremental, scalable aggregation of temporally evolving multimodal retinal images. Specifically, we introduce Incremental LoRA (I-LoRA) layers tailored to each modality, enabling the LLM to progressively incorporate modality-specific knowledge based on previously learned representations, thereby supporting dynamic integration of unbalanced modalities. In parallel, to provide well-perceived modality visual contexts for prompting LLM, we propose a Spatiotemporal-aware Expert (SAE) module that captures both spatial anatomical structures and temporal pattern dynamics within each modality, enhancing LLM's ability to perceive subtle, disease-relevant changes over time. By assigning SAE module with specific LoRA for incremental modality learning, MoEA-Net supports continual knowledge accumulation, enabling clinical inference even with incomplete modalities. The main contributions can be summarized as:

- We propose a novel Modality-Incremental Expert Aggregation Network, namely MoEA-Net, which unifies inter-modal integration and intra-modal perception in an incremental and scalable manner for boosting retinal prognostic prediction.
- We develop I-LoRA layers for progressive, modality-aware knowledge integration, and propose a Spatiotemporal-aware Expert module to capture fine-grained anatomical and temporal features.
- We comprehensively validate our approach on subretinal fluid change and visual recovery classification tasks. The results demonstrate that our model achieves SoTA performance.

2 Related Works

2.1 Modality-Aware Retinal Disease Diagnosis

OCT and fundus photography are well-established imaging modalities widely used in ophthalmology for retinal disease diagnosis (Ly et al. 2019; Greig, Duker, and Waheed 2020). OCT offers high-resolution cross-sectional views of retinal layers, while fundus images provide a wide-field, planar representation of retinal structures (Drexler et al. 2001). Driven by advances in deep learning for vision, language areas (Zhang et al. 2025b; Shi et al. 2023), ophthalmic diagnostics have also evolved significantly. Early approaches (Liefers et al. 2017; Li et al. 2021b; He et al. 2023; Hu et al. 2019) primarily leverage CNN-based, single-modality OCT analysis and transfer learning, achieving promising results for detecting AMD and diabetic retinopathy. More recently, multimodal learning frameworks have been explored to jointly leverage OCT and fundus images (He et al. 2021; Wang et al. 2022, 2024a). Wang et al. (2022) propose a dual-stream model for AMD classification that independently encodes each modality before feature fusion. However, multimodal methods are hindered by the scarcity of data, a problem rooted in clinical privacy. Although some approaches (Guo et al. 2024, 2025) aim to mitigate this, the fundamental lack of diverse datasets remains. And in clinical practice, strict pairing between fundus and OCT images is often unattainable, and heterogeneity between modalities further complicates model training. Unlike prior works, we explore the prognostic in a new modality OCTA, and target a more complex setting: temporal analysis of retinal images for prognostic prediction, which requires modeling both temporal and spatial correlations of modalities. By utilizing our modality incremental mechanism, the model can dynamically integrate new modalities (e.g., NVL), enhancing model scalability, adaptability, and clinical applicability.

2.2 Multi-modal Large Language Models

The emergence of MLLMs has significantly advanced medical visual understanding tasks (Zhang et al. 2025c; Zheng et al. 2024). These models, trained on large-scale image-text pairs and fine-tuned using prompt-based strategies, exhibit robust cross-modal alignment and reasoning abilities. In ophthalmology, recent efforts (Zhao et al. 2023; Deng et al. 2024) integrate textual prompts with imaging modalities to enhance knowledge representation and disease progression modeling. Wang et al. (2024b) further improves disease-specific understanding by aligning ophthalmic knowledge graphs with visual features. While these studies highlight the potential of MLLMs in capturing complex medical patterns, they overlook the inherent heterogeneity among ophthalmic modalities, which is an omission that limits their robustness in real-world settings. Additionally, modality imbalance also remains a challenge, reducing model flexibility and generalization. In contrast, our model employs the SAE module to distill anatomical changes over time, enabling the LLM to incrementally interpret multi-modal retinal patterns with improved temporal and spatial coherence.

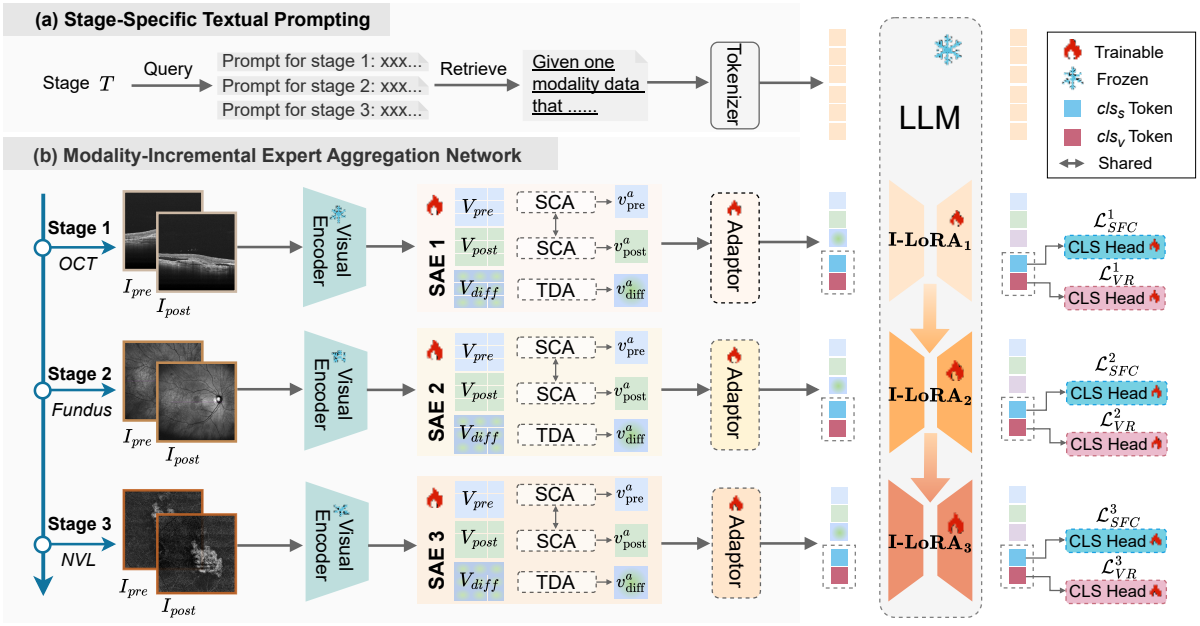

Figure 2: Illustration of the proposed MoEA-Net framework. (a) Stage-specific textual prompting constructs tailored prompts for each modality stage based on clinical context. (b) The Modality-Incremental Expert Aggregation Network integrates new modality-specific expert modules and I-LoRA branches in a progressive manner. Extracted features are projected into the LLM embedding space and injected into prompts for stage-wise clinical prediction.

3 Methods

3.1 Overall Framework

An overview of MoEA-Net is shown in Figure 2. Given the primary modality $OCT = \{I_{pre}, I_{post}\}$, the objective is multi-task prognostic prediction, including: (1) *Subretinal Fluid Change Classification (SFC)* for evaluating fluid dynamics, and (2) *Visual Recovery Classification (VR)* for binary assessment of visual function improvement.

Modality Pathological Learning To obtain disease-relevant visual representations from temporally paired OCT images, we employ BiomedCLIP (Zhang et al. 2023) as the visual encoder. Patch-level features are extracted as $V_{\text{pre}} = \{v_{\text{pre},1}, ..., v_{\text{pre},P}\} \in \mathbb{R}^{P\times d}$ and $V_{\text{post}} = \{v_{\text{post},1}, ..., v_{\text{post},P}\} \in \mathbb{R}^{P\times d}$ from the pre- and post-treatment images, where P denotes the number of image patches and d is the feature dimension.

To model both anatomical structure and temporal disease progression, we apply the proposed Spatiotemporal-Aware Expert (SAE) module to extract global semantic features:

$$V^a = SAE(V_{\text{pre}}, V_{\text{post}}), \tag{1}$$

where $V^a = \{v^a_{\text{pre}}, v^a_{\text{post}}, v^a_{\text{diff}}\} \in \mathbb{R}^{3\times d}$ encodes the overall condition before and after treatment, as well as the pathological difference between them. Subsequently, to interface with the LLM, we further employ a lightweight adaptor, realized as a two-layer multi-layer perceptron (MLP), which projects V^a into the word embedding space d_w within LLM. The resulting embeddings V_e are treated as modality-aware semantic tokens and concatenated directly into a multimodal prompt sequence.

Model Training We construct the multimodal input sequence $\mathcal{S}$ by concatenating the projected visual embeddings V_e, the textual prompt tokens T:

$$\mathcal{S} = [T, V_e, cls_s, cls_v], \tag{2}$$

where the cls_s and cls_v denote special classification tokens for SFC and VR tasks, respectively. This prompt is then decoded by the LLM, which is fine-tuned using Low-Rank Adaptation (LoRA) (Hu et al. 2022). The LoRA modules facilitate the absorption of disease-related semantics into the LLM while keeping most model parameters frozen.

Next, the final hidden states of the cls_s and cls_v tokens are leveraged for task-specific classification:

$$p_s = argmax(softmax(cls_s\theta_1)), \tag{3}$$

$$p_v = sigmoid(cls_v\theta_2), \tag{4}$$

where $\theta_1 \in \mathbb{R}^{d_w\times C}$ and $\theta_2 \in \mathbb{R}^{d_w\times 1}$ represent trainable classifiers for the two tasks, and C denotes the number of categories for SFC task.

The training objective combines a cross-entropy loss for SFC and a binary cross-entropy loss for VR:

$$\mathcal{L}_{\text{SFC}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(p_{i,c}), \tag{5}$$

$$\mathcal{L}_{\text{VR}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i\log(p_i)+(1-y_i)\log(1-p_i)\right], \tag{6}$$

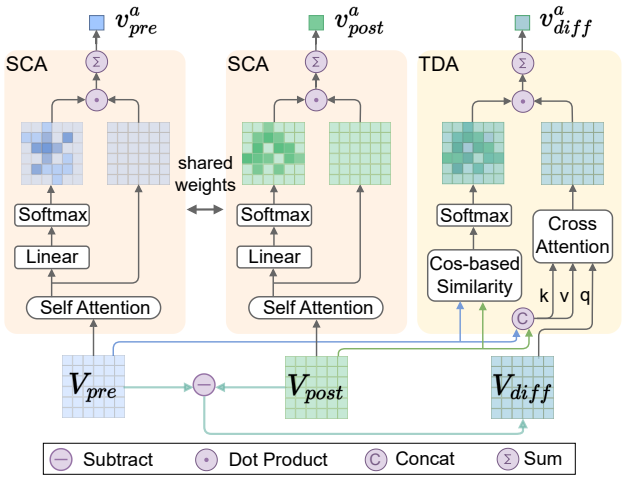

Figure 3: Illustration of the Spatiotemporal-Aware Expert (SAE) module. It consists of a Spatial Context Aggregation (SCA) branch with shared self-attention to enhance spatial features, and a Temporal Difference Aggregation (TDA) branch that models local changes via patch-wise subtraction and cross-attention. Both branches apply change-aware attention pooling to generate global descriptors.

where N is the batch size, $y_{i,c}$ and $p_{i,c}$ are the ground-truth label and predicted probability for class c in the fluid task, and y_i, p_i are the binary labels and predictions for visual acuity improvement. The final loss can be formulated as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{SFC}} + \lambda_2 \mathcal{L}_{\text{VR}}. \tag{7}$$

The loss coefficients λ_1 and λ_2 control the relative importance of each task.

3.2 Spatiotemporal-Aware Expert Module

To effectively model modality-specific pathological features, we propose a spatiotemporal-aware expert module with two functionally complementary submodules: a Spatial Context Aggregation (SCA) module and a Temporal Difference Aggregation (TDA) module. The former captures rich spatial contextual information by enhancing the spatial feature representations of both pre-treatment and post-treatment images through shared-weight encoding, enabling consistent perception of structural patterns across images. The latter is dedicated to modeling temporal differences, focusing on identifying disease-related changes over time by highlighting lesion-evolving regions while suppressing irrelevant background and stable tissues. As illustrated in Figure 3, the module integrates spatial and temporal cues through these two specialized pathways to comprehensively capture pathological variations.

Spatial Context Aggregation To enhance spatial understanding of retinal structures, we apply a shared multi-head self-attention (MHA) (Vaswani et al. 2017) to the pre- and post-treatment OCT features V_{pre} and V_{post}. MHA captures long-range dependencies among spatially distributed patches and yields refined contextual features:

$$V'_{\text{pre}} = MHA(V_{\text{pre}}), \quad V'_{\text{post}} = MHA(V_{\text{post}}), \tag{8}$$

where $V'_{\text{pre}}, V'_{\text{post}} \in \mathbb{R}^{P\times d}$ denote the enhanced spatial feature maps. Attention weights are shared across time to ensure consistent encoding.

Next, we employ a Diagnostic-aware Attention Pooling mechanism to generate compact global descriptors, denoted as DAP$(\cdot)$, which aggregates patch-level features based on their semantic importance. Specifically, for input feature map $V = \{v_1, ..., v_P\} \in \mathbb{R}^{P\times d}$, we compute the global token as:

$$\mathbf{w}_l = softmax(V\theta_l), \tag{9}$$

$$v^a = \sum_{i=1}^{P} w_i \cdot v_i, \tag{10}$$

where $\theta_l \in \mathbb{R}^{d\times 1}$ is a learnable projection vector, $\mathbf{w}_l = \{w_1, ..., w_P\} \in \mathbb{R}^P$ denotes the attention weights across spatial patches, and $v^a \in \mathbb{R}^d$ is the resulting global vector. This mechanism emphasizes lesion-relevant regions while suppressing background noise. We apply this operation to both pre- and post-treatment enhanced feature maps:

$$v^a_{\text{pre}} = DAP(V'_{\text{pre}}), \quad v^a_{\text{post}} = DAP(V'_{\text{post}}). \tag{11}$$

The resulting global descriptions, v^a_{pre} and v^a_{post}, are forwarded to downstream modules for further cross-temporal modeling and clinical prediction.

Temporal Difference Aggregation To explicitly capture disease-induced structural changes over time, we propose a Temporal Difference Aggregation (TDA) module, which focuses on modeling fine-grained temporal variations between pre- and post-treatment images. Given the spatially refined OCT features V_{pre} and V_{post}, we first compute the patch-wise difference:

$$V_{\text{diff}} = V_{\text{post}} - V_{\text{pre}}, \tag{12}$$

Then, each difference token $v_{\text{diff},i}$ serves as the query in a cross-attention mechanism, where the corresponding pre- and post-treatment features at position i act jointly as keys and values:

$$q = v_{\text{diff},i} W_q, \tag{13}$$

$$k = [v_{\text{pre},i}, v_{\text{post},i}] W_k, \tag{14}$$

$$v = [v_{\text{pre},i}, v_{\text{post},i}] W_v, \tag{15}$$

$$v'_{\text{diff},i} = Attn(q, k, v), \tag{16}$$

where $W_q, W_k, W_v \in \mathbb{R}^{d\times d}$ are learnable projection matrices, and $[\cdot,\cdot]$ denotes token concatenation. This formulation enables each temporal difference token to contextualize local changes using its spatial history.

Similarly, to obtain a global representation of disease progression, we introduce a Confidence-aware Weighted Pooling (CWP) mechanism. For each location, we compute the cosine similarity between $v_{\text{pre},i}$ and $v_{\text{post},i}$, assigning lower weights to highly similar (i.e., unchanged) regions:

$$z_i = -\cos(v_{\text{pre},i}, v_{\text{post},i}), \tag{17}$$

$$z'_i = \frac{\exp(z_i/\tau)}{\sum_{j=1}^{P} \exp(z_j/\tau)}, \tag{18}$$

$$v^a_{\text{diff}} = \sum_{i=1}^{P} v'_{\text{diff},i} \cdot z'_i, \tag{19}$$

Methods	Subretinal Fluid Change (SFC)				Visual Recovery (VR)			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
(a) Traditional Methods based on Deep Learning								
VGG-16 (Simonyan and Zisserman 2015)†	0.614	0.453	0.507	0.469	0.517	0.510	0.518	0.514
GoogleNet (Szegedy et al. 2015)†	0.607	0.451	0.519	0.467	0.521	0.514	0.557	0.534
ResNet-152 (He et al. 2016)†	0.659	0.435	0.427	0.430	0.538	0.530	0.557	0.543
DenseNet-121 (Huang et al. 2017)†	0.634	0.404	0.406	0.403	0.508	0.501	0.585	0.540
(b) Multimodal Medical Data Fusion or Distillation Methods								
MM-MIL (Li et al. 2021a)†	0.650	0.466	0.481	0.473	0.524	0.513	0.677	0.584
MSAN (He et al. 2021)†	0.675	0.491	0.526	0.490	0.525	0.515	0.640	0.571
MM-CNN (Wang et al. 2022)†	0.673	0.519	0.561	0.505	0.534	0.527	0.531	0.529
FDDM (Wang et al. 2023)†	0.679	0.518	0.596	0.535	0.521	0.511	0.677	0.583
GeCoM-Net (Wang et al. 2024a)†	0.684	0.497	0.520	0.489	0.540	0.534	0.531	0.533
OCT-CoDA (Wang et al. 2025)†	0.681	0.582	0.574	0.502	0.541	0.530	0.608	0.566
(c) Multimodal Large Language Models								
CD-Chat (Noman et al. 2024)†	0.709	0.544	0.596	0.563	0.563	0.545	0.695	0.611
Baseline	0.708	0.544	0.607	0.561	0.536	0.522	0.705	0.600
MoEA-Net (Ours)	**0.752**	**0.602**	**0.639**	**0.617**	**0.583**	**0.560**	**0.722**	**0.631**

Table 1: Performance comparison of various methods on AMD-OCTA for Subretinal Fluid Change and Visual Recovery tasks. † denotes results reproduced based on the authors' released code or descriptions.

where $\tau > 0$ is a temperature scale that controls the strength to soften the distribution. A lower τ emphasizes patches with greater change, improving sensitivity to critical progression regions.

The resulting global embedding $v^a_{\text{diff}} \in \mathbb{R}^d$ serves as a compact summary of structural differences, which is passed to downstream modules for joint representation learning and clinical prediction.

3.3 Progressive Knowledge Incremental Learning

To address the issue of *modality imbalance* in clinical multimodal data, we propose a progressive, stage-wise knowledge accumulation mechanism, termed **Incremental LoRA (I-LoRA)**. This design enables continual integration of new modalities and corresponding expert modules *without full model retraining*, effectively overcoming challenges such as catastrophic forgetting (Kirkpatrick et al. 2016), poor generalization, and high computational cost.

As illustrated in Figure 2, after each training stage, the corresponding *expert module* and *I-LoRA branch* (e.g., SAE_1 + I-LoRA_1 for OCT) are *frozen*, then newly introduced modalities are processed by their modality-tailored experts and I-LoRA branches. This modular training strategy enables the model to *preserve previously acquired knowledge while continuously integrating new modality-specific information*. All extracted features are fused into multimodal prompts and fed into a pretrained large language model (LLM) for end-to-end optimization. At each training stage t, the LLM receives a stage-tailored prompt consisting of three components: the stage-specific textual prompt $T^{(t)}$, the projected visual embeddings from the new modality $\mathcal{U}(\mathcal{I}^{(t)})$, and two special classification tokens cls_s and cls_v. The output representation of the LLM at stage t is denoted as:

$$\mathcal{R}^{(t)} = \mathcal{M}^{(t)}_{\text{LLM}}\left(\left[T^{(t)}, \mathcal{U}(\mathcal{I}^{(t)}), cls_s, cls_v\right]\right), \tag{20}$$

where $\mathcal{R}^{(t)}$ is used for downstream clinical predictions.

The underlying LLM architecture is progressively updated through stage-wise injection of I-LoRA modules. After completing training at stage t, the newly trained LoRA parameters $\theta^{(t)} = \{A^{(t)}, B^{(t)}\}$ are frozen and integrated into the model:

$$\mathcal{M}^{(t)}_{\text{LLM}} = \mathcal{M}^{(t-1)}_{\text{LLM}} + \theta^{(t)}, \tag{21}$$

where $A^{(t)} \in \mathbb{R}^{d_{\text{in}} \times r}$ and $B^{(t)} \in \mathbb{R}^{r \times d_{\text{out}}}$ are low-rank adaptation matrices, and r is the adaptation rank. This incremental update enables the model to adapt to new modalities efficiently, without altering previously learned parameters.

Overall, this progressive adaptation strategy enables flexible and memory-efficient integration of new modalities, with each LoRA branch encapsulating modality-specific knowledge that is incrementally merged into the LLM. It avoids catastrophic forgetting and supports scalable deployment.

4 Experiments and Results

4.1 Datasets, Metrics and Settings

Datasets We construct the AMD-OCTA dataset, a longitudinal retinal dataset for this study, comprising 23,242 multimodal images from 5,903 anonymized AMD samples. Each sample includes data from two time points, with the following modalities: (1) OCT, (2) Fundus, (3) NVL, (4) LMCV-Pseudo, and (5) LMCV. The samples are labeled for the SFC task based on subretinal fluid absorption status (no absorption, partial absorption, and complete absorption) and for the VR task based on visual acuity status (improved or unchanged). Following expert recommendations, we excluded

Method	Subretinal Fluid Change(SFC)			
	Accuracy	Precision	Recall	F1
VGG-16	0.594	0.338	0.323	0.329
GoogleNet	0.623	0.414	0.391	0.386
ResNet-152	0.642	0.423	0.421	0.415
DenseNet-121	0.607	0.369	0.367	0.367
MM-MIL	0.619	0.485	0.342	0.351
MSAN	0.645	0.412	0.436	0.422
MM-CNN	0.654	0.445	0.398	0.415
FDDM	0.670	0.450	0.406	0.423
GeCoM-Net	0.667	0.315	0.355	0.334
OCT-CoDA	0.676	0.507	0.409	0.435
CD-Chat	0.692	0.522	0.445	0.468
Baseline	0.698	0.562	0.430	0.467
MoEA-Net (Ours)	**0.714**	**0.593**	**0.459**	**0.506**

Table 2: Comparison of different methods on OCT4DME for predicting Subretinal Fluid Structural Change.

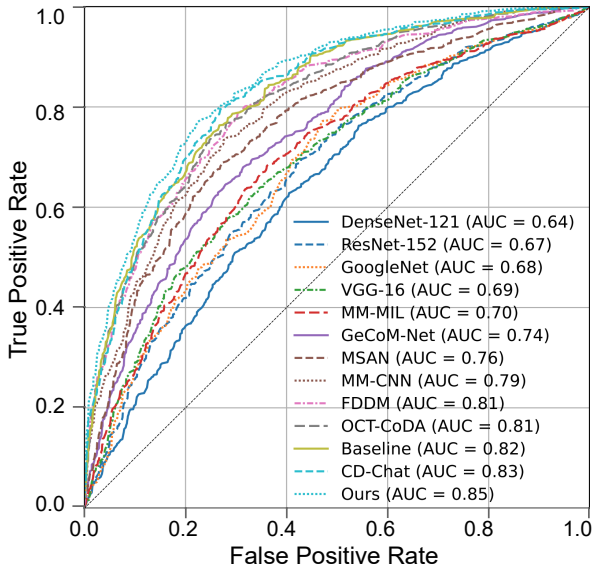

Figure 4: ROC curve visualizations for subretinal fluid change classification performance on the AMD-OCTA dataset, with AUC scores for different methods.

the clinically irrelevant LMCV-Pseudo and LMCV modalities from model training. Further details and examples are provided in Appendix A.1.

To evaluate the generalization of our model, we also perform experiments on the open-sourced OCT4DME dataset (Zhang et al. 2025a), which contains 2,503 DME samples and 9,976 multimodal images, with only OCT and Fundus modalities. Similar to our AMD-OCTA dataset, OCT4DME contains images from two time points and is annotated for the occurrence of subretinal fluid in the SFC task, with the following categories: no subretinal fluid, resolution of subretinal fluid, onset of subretinal fluid, and persistent subretinal fluid. Both datasets were split into training, validation, and test sets in a 7:1:2 ratio.

Metrics We comprehensively evaluate the model performance using standard classification metrics: Accuracy, Precision, Recall, F1-score, and the Area Under the ROC Curve (AUC) (Litjens et al. 2017; Esteva et al. 2017). These metrics are well-suited to capture the model's ability to track temporal changes in subretinal fluid change and visual recovery status.

Settings We resize all modality images to 224×224 pixels and extract features using the vision encoder of BiomedCLIP (Zhang et al. 2023). We adopt LLaMA3-8B (Meta 2024) as the LLM backbone, which is integrated with the BiomedCLIP's vision encoder to construct our baseline model. The model is trained in a progressive multi-stage manner: 60 epochs in Stage 1, followed by 50 epochs in Stage 2, and 40 epochs in Stage 3. We set loss coefficients $\lambda_1 = 1$ and $\lambda_2 = 2$, selected based on parameter sensitivity experiments (see Appendix A.2). The temperature parameter τ is set to 0.6. Following prior work (Zhao et al. 2021), we apply data augmentations including random rotation, scaling, and flipping to improve generalization (see Appendix A.3 for details). All comparative methods utilize the same augmentation pipeline to ensure fairness. More experimental details can be found in Appendix A.4.

4.2 Comparison Studies

We conduct a comprehensive comparison against state-of-the-art methods across three categories: (1) traditional single-modal methods, (2) multimodal fusion-based methods, and (3) MLLM-based methods, as summarized in Table 1. Evaluation is performed on both the SFC and VR tasks.

Our model achieves the highest overall performance across all metrics. Traditional deep learning models, such as ResNet-152 and DenseNet-121, yield moderate gains over GoogleNet (e.g., $\sim$ 3.0% $\uparrow$ Accuracy gain on SFC task), yet still underperform in key metrics like F1 score and Recall. In contrast, multimodal fusion and knowledge distillation approaches, e.g., GeCoM-Net and OCT-CoDA, achieve stronger results across several metrics (e.g., OCT-CoDA: 58.2% Precision on SFC, +17.8% over DenseNet), highlighting the benefits of multimodal learning. However, their F1 scores remain limited, only $\leq$ 53.5% on SFC and $\leq$ 58.4% on VR. These results diverge from their performance on original tasks with single-timepoint inputs, underscoring the increased complexity of our temporal learning setting. The MLLM-based model CD-Chat achieves F1 scores of 56.3% (SFC) and 61.1% (VR), though the gains primarily stem from its strong backbone rather than effective cross-modal fusion. By contrast, our model achieves substantial improvements on both tasks, with 75.2% Accuracy on SFC (+4.3% than CD-Chat) and 72.2% recall on VR (+2.7%). Besides, we also visualize the ROC curves of different methods, as shown in Figure 4, where our MoEA-Net achieves the highest AUC.

To further demonstrate the advantages of our framework, we applied all methods to the open-sourced OCT4DME

Methods	Modalities			Experts		Subretinal Fluid Change (SFC)					Visual Recovery (VR)				
	OCT	Fundus	NVL	SCA	TDA	Acc	Prec	Rec	F1	AUC	Acc	Prec	Rec	F1	AUC
Baseline	✓					0.708	0.544	0.607	0.561	0.824	0.536	0.522	0.705	0.600	0.566
(a)	✓			✓		0.713	0.560	0.631	0.578	0.793	0.538	0.502	0.671	0.574	0.586
(b)	✓				✓	0.716	0.557	0.624	0.577	0.839	0.542	0.505	0.624	0.558	0.559
(c)	✓			✓	✓	0.720	0.566	0.634	0.591	0.831	0.565	0.524	0.621	0.569	0.579
(d)	✓	✓		✓	✓	0.737	0.572	0.630	0.593	0.835	0.573	0.536	0.613	0.572	0.586
Ours	✓	✓	✓	✓	✓	**0.752**	**0.602**	**0.639**	**0.617**	**0.846**	**0.583**	**0.560**	**0.722**	**0.631**	**0.608**

Table 3: Ablation study on the effect of modality combinations and expert module settings across two retinal prognostic prediction tasks: subretinal fluid absorption classification and visual recovery classification.

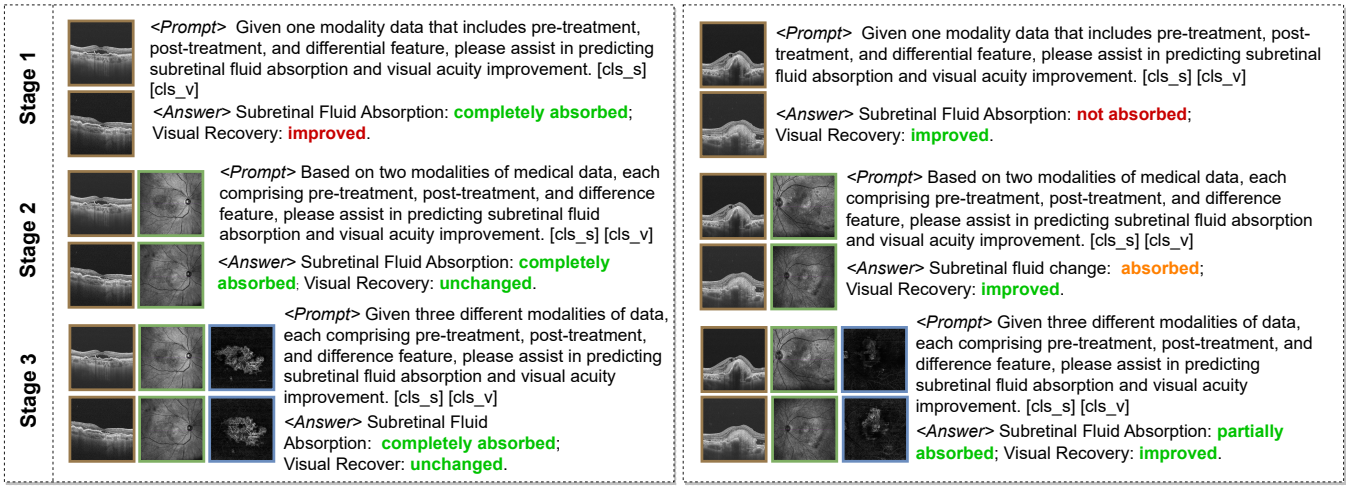

Figure 5: Visualization of prediction results for two cases across different model stages. Green indicates correct predictions, Red denotes incorrect predictions, and Orange highlights predictions that are close to the ground truth.

dataset. As shown in Table 2, our method achieves 71.4% Accuracy and 59.3% Precision, outperforming all the compared methods. Notably, this performance is achieved under the severe label imbalance in the OCT4DME dataset. These results provide evidence of our model's effectiveness in modality-incremental learning and spatiotemporal modeling.

4.3 Ablation Studies

To evaluate the contributions of expert modules and modality integration, we conduct ablation studies in Table 3. The baseline model utilizes only the OCT modality without any expert modules, achieving moderate performance (e.g., 56.1% F1 on the SFC task). With the inclusion of the SCA module (a) or the TDA module (b), the model achieves higher F1 ($\sim$ 2.0% $\uparrow$) and recall ($\sim$ 2.0% $\uparrow$) scores on the SFC task, indicating that these modules can effectively aggregate spatial and temporal clues, leading to a better retinal prognostic prediction. Combining both SCA and TDA modules, (c) achieves further gains (72.0% Accuracy on SFC), highlighting their complementary effects. Comparisons among settings (c), (d), and *Ours* also demonstrate the advantages of modality integration, which benefits both tasks with enhanced visual contexts.

4.4 Qualitative Analysis

To further illustrate the effectiveness of our MoEA-Net, we visualize the model's predictions across different stages in Figure 5. At stage 1, where only OCT input is used, the model produces erroneous outputs (e.g., misclassifying visual recovery as "improved"). As additional modalities are incorporated in stages 2 and 3, prediction accuracy improves noticeably on both tasks, highlighting the benefit of our incremental modality learning framework in capturing complementary information.

5 Conclusion

We propose MoEA-Net, a Modality-Incremental Expert Aggregation Network for retinal prognostic prediction. MoEA-Net unifies inter-modal integration and intra-modal perception in multi-modal temporal retinal imaging by progressively combining the Spatiotemporal-aware Expert module and incremental LoRA layers to LLM, which supports continual knowledge accumulation and improves accurate reasoning. Experimental results show that MoEA-Net achieves SoTA performance on both subretinal fluid change and visual recovery classification. We believe this study offers valuable insights into retinal visual learning and contributes to more generalizable computer-aided diagnosis.

Acknowledgments

This work was supported in part by the National High-Level Hospital Clinical Research Funding under Grant BJ-2024-089, and in part by the Beijing Natural Science Foundation under Grant L254040.

References

Deng, Z.; Gao, W.; Chen, C.; Niu, Z.; Gong, Z.; Zhang, R.; Cao, Z.; Li, F.; Ma, Z.; Wei, W.; and Ma, L. 2024. OphGLM: An ophthalmology large language-and-vision assistant. *Artif. Intell. Medicine*, 157: 103001.

Drexler, W.; Morgner, U.; Ghanta, R. K.; Kärtner, F. X.; Schuman, J. S.; and Fujimoto, J. G. 2001. Ultrahigh-resolution ophthalmic optical coherence tomography. *Nature medicine*, 7(4): 502–507.

Esteva, A.; Kuprel, B.; Novoa, R. A.; Ko, J.; Swetter, S. M.; Blau, H. M.; and Thrun, S. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nat.*, 542(7639): 115–118.

Greig, E. C.; Duker, J. S.; and Waheed, N. K. 2020. A practical guide to optical coherence tomography angiography interpretation. *International journal of retina and vitreous*, 6: 1–17.

Guo, P.; Wang, R.; Zeng, S.; Zhu, J.; Jiang, H.; Wang, Y.; Zhou, Y.; Wang, F.; Xiong, H.; and Qu, L. 2025. Exploring the vulnerabilities of federated learning: A deep dive into gradient inversion attacks. *arXiv preprint arXiv:2503.11514*.

Guo, P.; Zeng, S.; Wang, Y.; Fan, H.; Wang, F.; and Qu, L. 2024. Selective aggregation for low-rank adaptation in federated learning. *arXiv preprint arXiv:2410.01463*.

Hao, Y.; Liu, S.; and Yu, Z. 2022. Value of Combining Optical Coherence Tomography with Fundus Photography in Screening Retinopathy in Patients with High Myopia. *Journal of Healthcare Engineering*, 2022.

He, J.; Wang, J.; Han, Z.; Ma, J.; Wang, C.; and Qi, M. 2023. An interpretable transformer network for the retinal disease classification using optical coherence tomography. *Scientific Reports*, 13(1): 3637.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778. IEEE Computer Society.

He, X.; Deng, Y.; Fang, L.; and Peng, Q. 2021. Multi-Modal Retinal Image Classification With Modality-Specific Attention Network. *IEEE Trans. Medical Imaging*, 40(6): 1591–1602.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Hu, J.; Chen, Y.; Zhong, J.; Ju, R.; and Yi, Z. 2019. Automated Analysis for Retinopathy of Prematurity by Deep Neural Networks. *IEEE Trans. Medical Imaging*, 38(1): 269–279.

Huang, G.; Liu, Z.; van der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 2261–2269. IEEE Computer Society.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N. C.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; Hassabis, D.; Clopath, C.; Kumaran, D.; and Hadsell, R. 2016. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796.

Li, X.; Zhou, Y.; Wang, J.; Lin, H.; Zhao, J.; Ding, D.; Yu, W.; and Chen, Y. 2021a. Multi-Modal Multi-Instance Learning for Retinal Disease Recognition. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, 2474–2482. ACM.

Li, Y.; Zhang, Y.; Wang, Y.; Wang, Y.; Li, Y.; and Chen, Y. 2021b. Optical coherence tomography-based short-term effect prediction of anti-vascular endothelial growth factor treatment in neovascular age-related macular degeneration using sensitive structure guided network. *Computers in Biology and Medicine*, 135: 104607.

Liefers, B.; Venhuizen, F. G.; Schreur, V.; van Ginneken, B.; Hoyng, C.; Fauser, S.; Theelen, T.; and Sánchez, C. I. 2017. Automatic detection of the foveal center in optical coherence tomography. *Biomedical Optics Express*, 8(11): 5160–5178.

Litjens, G.; Kooi, T.; Bejnordi, B. E.; Setio, A. A. A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J. A. W. M.; van Ginneken, B.; and Sánchez, C. I. 2017. A survey on deep learning in medical image analysis. *Medical Image Anal.*, 42: 60–88.

Ly, A.; Phu, J.; Katalinic, P.; and Kalloniatis, M. 2019. An evidence-based approach to the routine use of optical coherence tomography. *Clinical and Experimental Optometry*, 102(3): 242–259.

Meta, A. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Noman, M.; Ahsan, N.; Naseer, M.; Cholakkal, H.; Anwer, R. M.; Khan, S.; and Khan, F. S. 2024. Cdchat: A large multimodal model for remote sensing change description. *arXiv preprint arXiv:2409.16261*.

Shafiq, M.; Fan, Q.; Alghamedy, F. H.; and Obidallah, W. J. 2024. DualEye-FeatureNet: A Dual-Stream Feature Transfer Framework for Multi-Modal Ophthalmic Image Classification. *IEEE Access*, 12: 143985–144008.

Shi, Y.; Ji, J.; Zhang, X.; Qu, L.; and Liu, Y. 2023. Granularity Matters: Pathological Graph-driven Cross-modal Alignment for Brain CT Report Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 6617–6630.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S. E.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich,

A. 2015. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, 1–9. IEEE Computer Society.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, L.; Dai, W.; Jin, M.; Ou, C.; and Li, X. 2023. Fundus-Enhanced Disease-Aware Distillation Model for Retinal Disease Classification from OCT Images. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2023 - 26th International Conference, Vancouver, BC, Canada, October 8-12, 2023, Proceedings, Part VII*, volume 14226 of *Lecture Notes in Computer Science*, 639–648. Springer.

Wang, L.; Qi, C.; Ou, C.; An, L.; Jin, M.; Kong, X.; and Li, X. 2025. MultiEYE: Dataset and Benchmark for OCT-Enhanced Retinal Disease Recognition From Fundus Images. *IEEE Trans. Medical Imaging*, 44(4): 1711–1722.

Wang, W.; Li, X.; Xu, Z.; Yu, W.; Zhao, J.; Ding, D.; and Chen, Y. 2022. Learning Two-Stream CNN for Multi-Modal Age-Related Macular Degeneration Categorization. *IEEE J. Biomed. Health Informatics*, 26(8): 4111–4122.

Wang, Y.; Zhen, L.; Tan, T.; Fu, H.; Feng, Y.; Wang, Z.; Xu, X.; Goh, R. S. M.; Ng, Y.; Calhoun, C.; Tan, G. S. W.; Sun, J. K.; Liu, Y.; and Ting, D. S. W. 2024a. Geometric Correspondence-Based Multimodal Learning for Ophthalmic Image Analysis. *IEEE Trans. Medical Imaging*, 43(5): 1945–1957.

Wang, Z.; Jiang, X.; Gao, C.; Dong, F.; Dai, W.; Wang, B.; Yan, B.; Chen, Q.; Huang, W.; Zhang, T.; and Chen, Y. 2024b. EyeGraphGPT: Knowledge Graph Enhanced Multimodal Large Language Model for Ophthalmic Report Generation. In *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2024*, 3784–3789.

Watanabe, T.; Hiratsuka, Y.; Kita, Y.; Tamura, H.; Kawasaki, R.; Yokoyama, T.; Kawashima, M.; Nakano, T.; and Yamada, M. 2022. Combining Optical Coherence Tomography and Fundus Photography to Improve Glaucoma Screening. *Diagnostics*, 12.

Zhang, H.; Wang, H.; Ge, S.; et al. 2023. BiomedCLIP: Contrastive Language-Image Pre-training for Biomedical Vision-Language Tasks. *arXiv preprint arXiv:2302.12556*.

Zhang, W.; Chotcomwongse, P.; Li, Y.; Xu, P.; Yao, R.; Zhou, L.; Zhou, Y.; Feng, H.; Zhou, Q.; Wang, X.; et al. 2025a. Predicting Diabetic Macular Edema Treatment Responses Using OCT: Dataset and Methods of APTOS Competition. *arXiv preprint arXiv:2505.05768*.

Zhang, X.; Jia, A.; Ji, J.; Qu, L.; and Ye, Q. 2025b. Intra- and Inter-Head Orthogonal Attention for Image Captioning. *IEEE Transactions on Image Processing*.

Zhang, X.; Shi, Y.; Ji, J.; Zheng, C.; and Qu, L. 2025c. MEP-Net: Medical Entity-Balanced Prompting Network for Brain CT Report Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25940–25948.

Zhao, H.; Ling, Q.; Pan, Y.; Zhong, T.; Hu, J.-Y.; Yao, J.; Xiao, F.; Xiao, Z.; Zhang, Y.; Xu, S.-H.; Wu, S.-N.; Kang, M.; Wu, Z.; Liu, Z.; Jiang, X.; Liu, T.; and Shao, Y. 2023. Ophtha-LLaMA2: A Large Language Model for Ophthalmology. arXiv:2312.04906.

Zhao, X.; Zhang, X.; Lv, B.; Meng, L.; Zhang, C.; Liu, Y.; and Chen, Y. 2021. Optical coherence tomography-based short-term effect prediction of anti-vascular endothelial growth factor treatment in neovascular age-related macular degeneration using sensitive structure guided network. *Graefe's Archive for Clinical and Experimental Ophthalmology*, 259(11): 3261–3269.

Zheng, C.; Ji, J.; Shi, Y.; Zhang, X.; and Qu, L. 2024. See Detail Say Clear: Towards Brain CT Report Generation via Pathological Clue-driven Representation Learning. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP*, 16542–16552. Association for Computational Linguistics.