Incomplete Multimodal Learning for Visual Acuity Prediction after Cataract Surgery Using Masked Self-Attention

Qian Zhou, Hua Zou*, Haifeng Jiang, and Yong Wang
Wuhan University, Wuhan, China
zouhua@whu.edu.cn

1. INTRODUCTION

While cataract surgery is routine, accurately predicting post-surgery visual acuity (BCVA) remains elusive. Conventional single-modal methods often disappoint. Our pioneering framework integrates preoperative images and patient demographics for a deeper multimodal understanding. Furthermore, we address incomplete data scenarios with a robust masked self-attention mechanism. Our approach outperforms state-of-the-art methods on the collected dataset, achieving a remarkable mean absolute error of 0.122 logMAR, with 94.3% of prediction errors within ±0.10 logMAR.

2. METHOD

2.1 Framework

Our framework (shown in Figure 1) consists of three parts: modality-specific encoder, multimodal fusion network, and BCVA prediction head. We use pre-trained Transformers—ViT for images and CLIP for text. A cross-modal Transformer combines features from different modalities with an attentional mask for missing data. BCVA prediction is performed using a fully connected (FC) layer.

2.2 Text-Encoder

We employ a pre-trained CLIP model as our text encoder. To enhance compatibility, physiological data is transformed into standardized sentences. For example, "male, 67 years old, preoperative visual acuity 0.52 logMAR" becomes "A 67-year-old male patient with preoperative visual acuity of 0.52 logMAR." This approach improves data consistency and simplifies semantic information extraction for the model.

2.3 Image-Encoder

ViT is used as the image encoder. Since each image in our dataset is associated with diagnostic keywords provided by ophthalmologists, we introduce an auxiliary classification loss within the image encoder. For each input image, a multi-label classification network is integrated after the image encoder to predict the diseases present in the image.

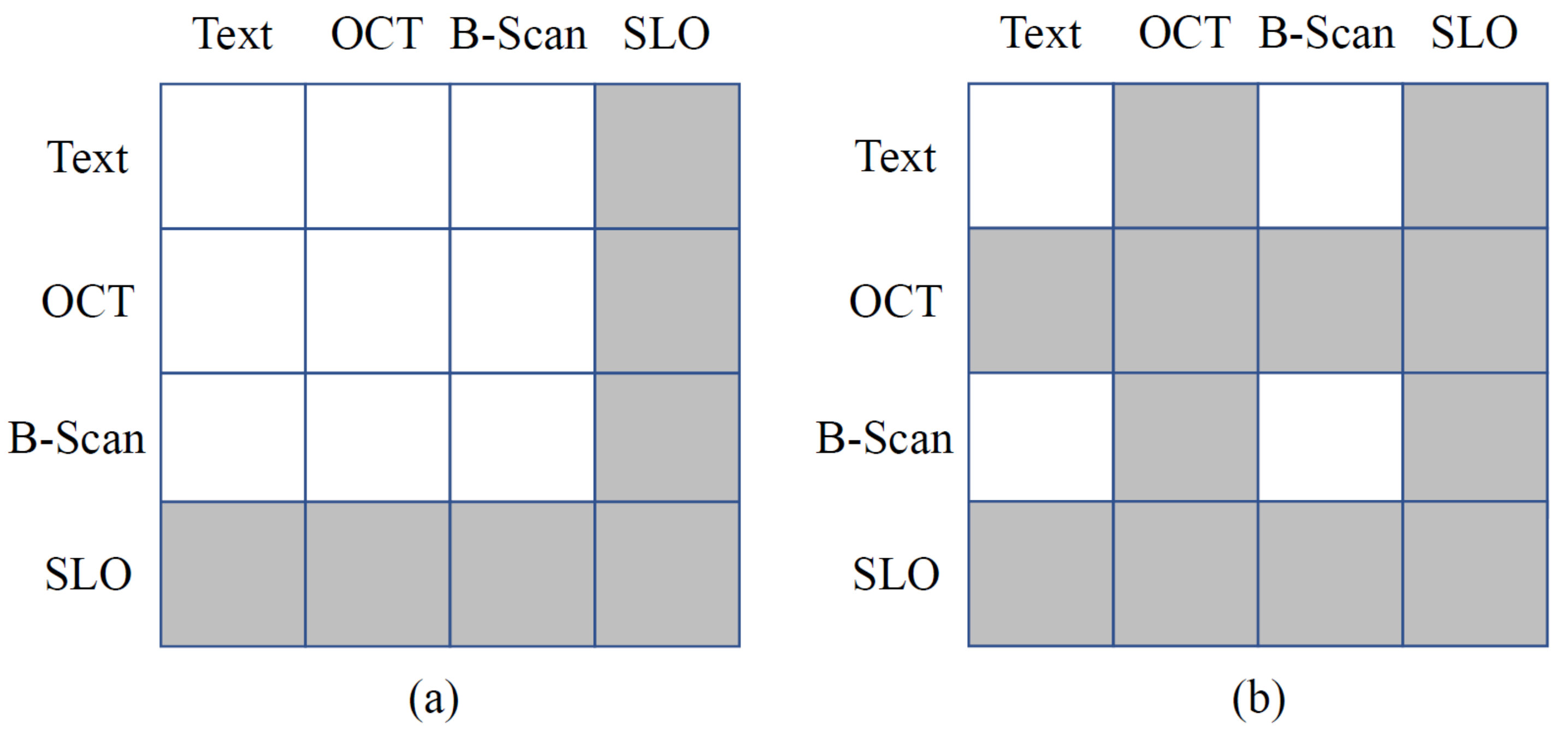
$$L_{CLS} = \sum W^i L^i_{BCE}$$

$$L = L_{MSE} + \alpha L_{CLS}$$

Here, W_i equals 1 if the i-th modality is available and 0 otherwise, α is a hyperparameter set to 0.5, L is training loss of the whole model.

2.4 Masked Self-Attention

Not all cases provide complete modalities (OCT, SLO, and Ultrasound). Representing missing modalities as 0 values can introduce noise to the model. To prevent this, we employ attentional masks within vanilla self-attention to exclude interactions between missing and available modalities. Note that the masked self-attention can be applied to both complete and incomplete multimodal fusion.



	Text	OCT	B-Scan	SLO
Text				▨
OCT				▨
B-Scan				▨
SLO	▨	▨	▨	▨

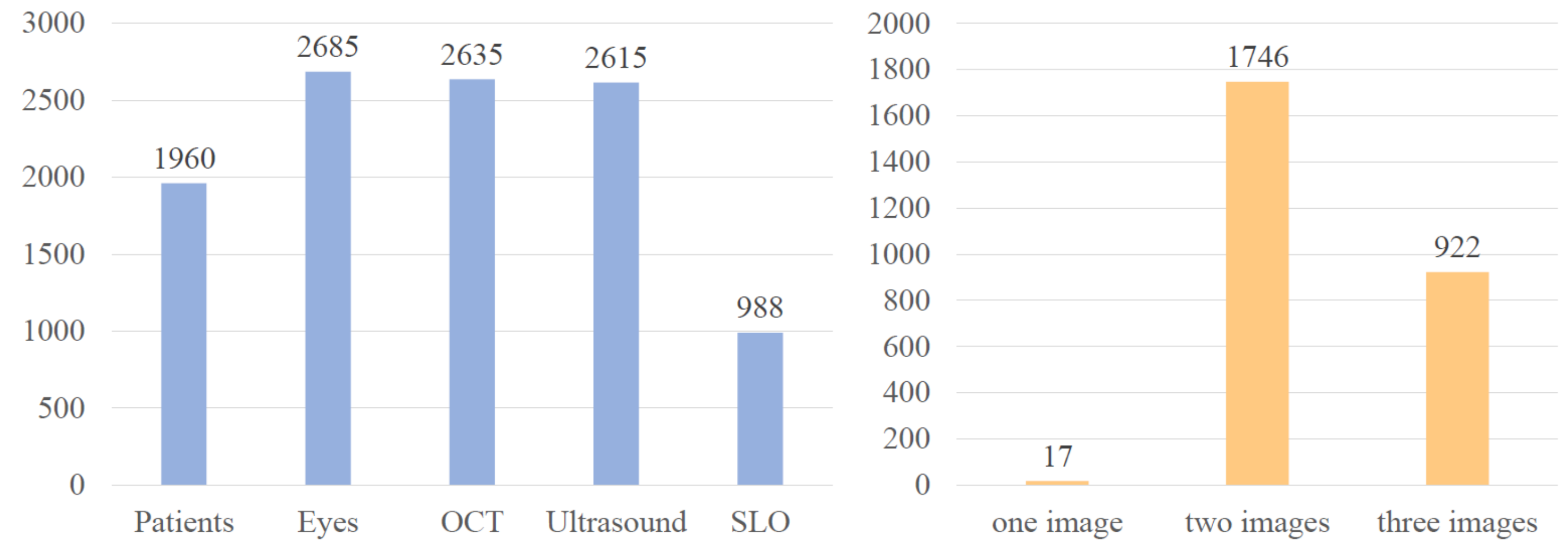
(a)

	Text	OCT	B-Scan	SLO
Text		▨		▨
OCT	▨	▨	▨	▨
B-Scan		▨		▨
SLO	▨	▨	▨	▨

(b)

Figure 2. An example of the attentional mask. (a) means only SLO is missing; (b) represents both OCT and SLO are missing.

2.5 Collected Dataset



(a) Numbers of each category in the collected dataset. (b) Numbers of eyes with different numbers of images.

Figure 3. Statistics for the collected dataset. (a) Number of patients, eyes, and three image modalities. (b) Number of multimodal or monomodal samples. Only one-third of cases have complete multimodal images.

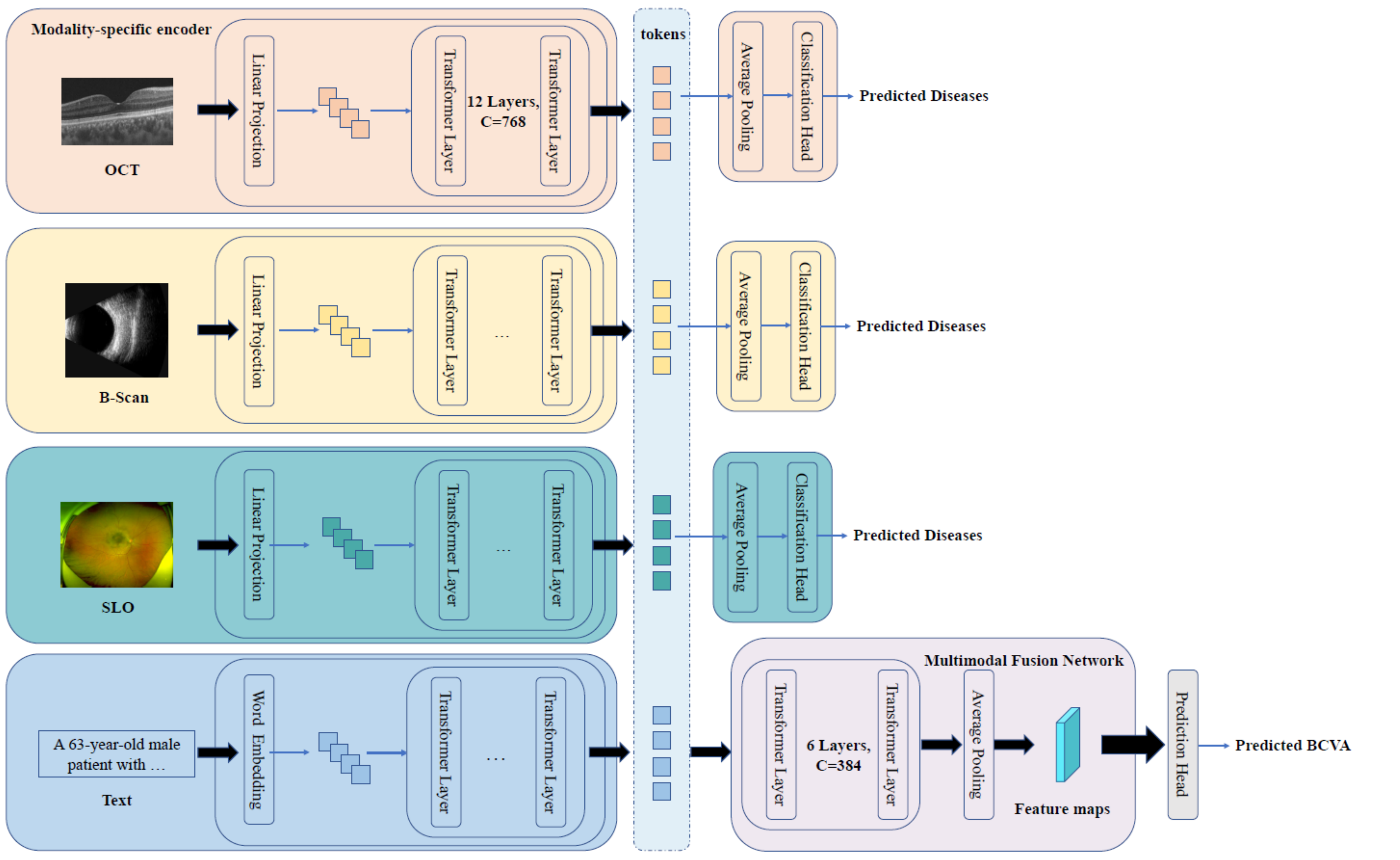


Figure 1. Pipeline of the proposed framework. The modality-specific encoders utilize vanilla multi-head self-attention. In contrast, the multimodal fusion network employs masked multi-head self-attention.

3. EVALUATION

3.1 Quantitative Performance

Dataset	Methods	MAE (↓)	SMAPE (↓)	Accuracy (↑)
Complete	CTT-Net [15] (OCT+Text)	0.168 ± 0.014	85.236 ± 3.277	0.887 ± 0.022
	CTT-Net [15] (OCT)	0.174 ± 0.013	89.635 ± 2.881	0.872 ± 0.016
	Wei *et al.* [16] (OCT)	0.237 ± 0.093	93.587 ± 3.236	0.723 ± 0.056
	Ours (OCT)	0.153 ± 0.012	65.615 ± 1.690	0.901 ± 0.018
	Ours (OCT + Text)	0.142 ± 0.009	62.550 ± 1.668	0.923 ± 0.014
Incomplete	Huang *et al.* [6]	0.176 ± 0.054	88.672 ± 3.051	0.854 ± 0.017
	Ma *et al.* [7]	0.139 ± 0.013	61.722 ± 2.007	0.917 ± 0.015
	Zhao *et al.* [19]	0.133 ± 0.021	59.673 ± 2.362	0.921 ± 0.021
	Ours	$\mathbf{0.122 \pm 0.007}$	$\mathbf{57.165 \pm 1.610}$	$\mathbf{0.943 \pm 0.012}$

3.2 Quantitative Performance

Figure 4 shows that the predicted and actual visual acuity means are quite close, with substantial overlap in the histograms. This suggests that the proposed method accurately predicts the majority of test samples. With the proposed method, the model can pay more attention to the most important foveal area in the fundus structures as shown in Figure 5.

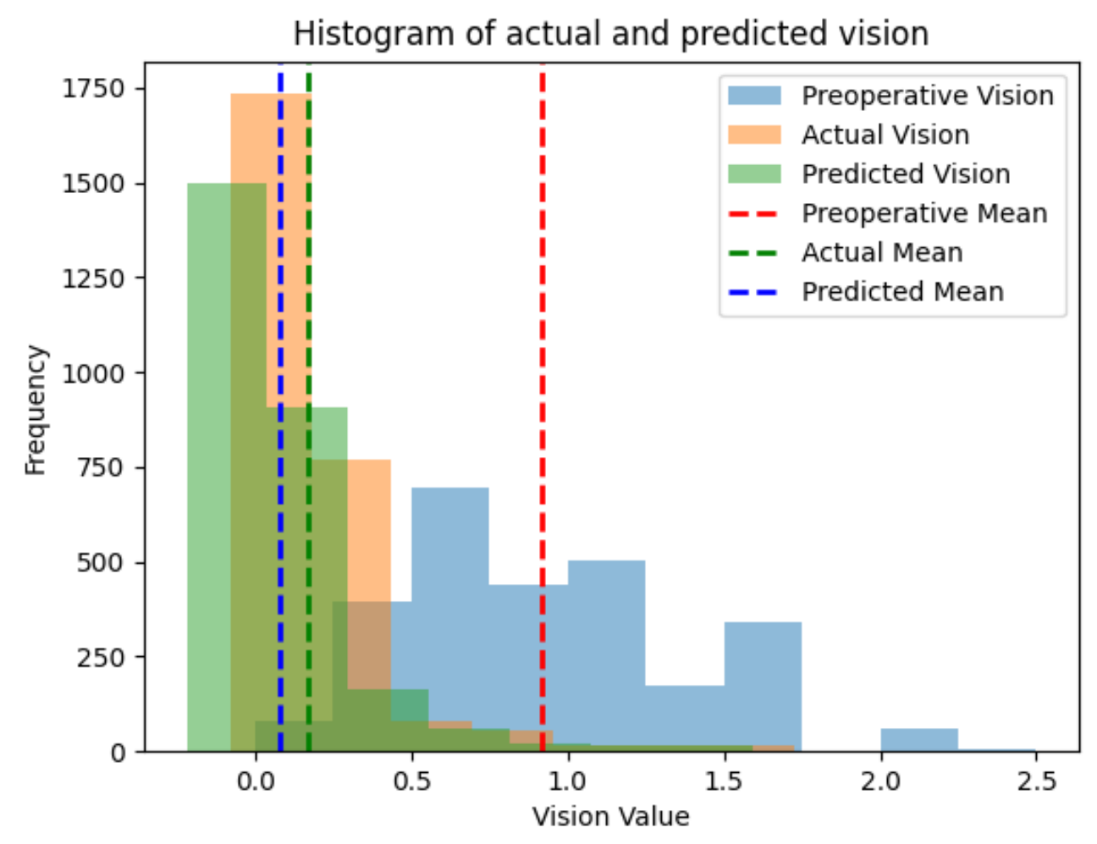
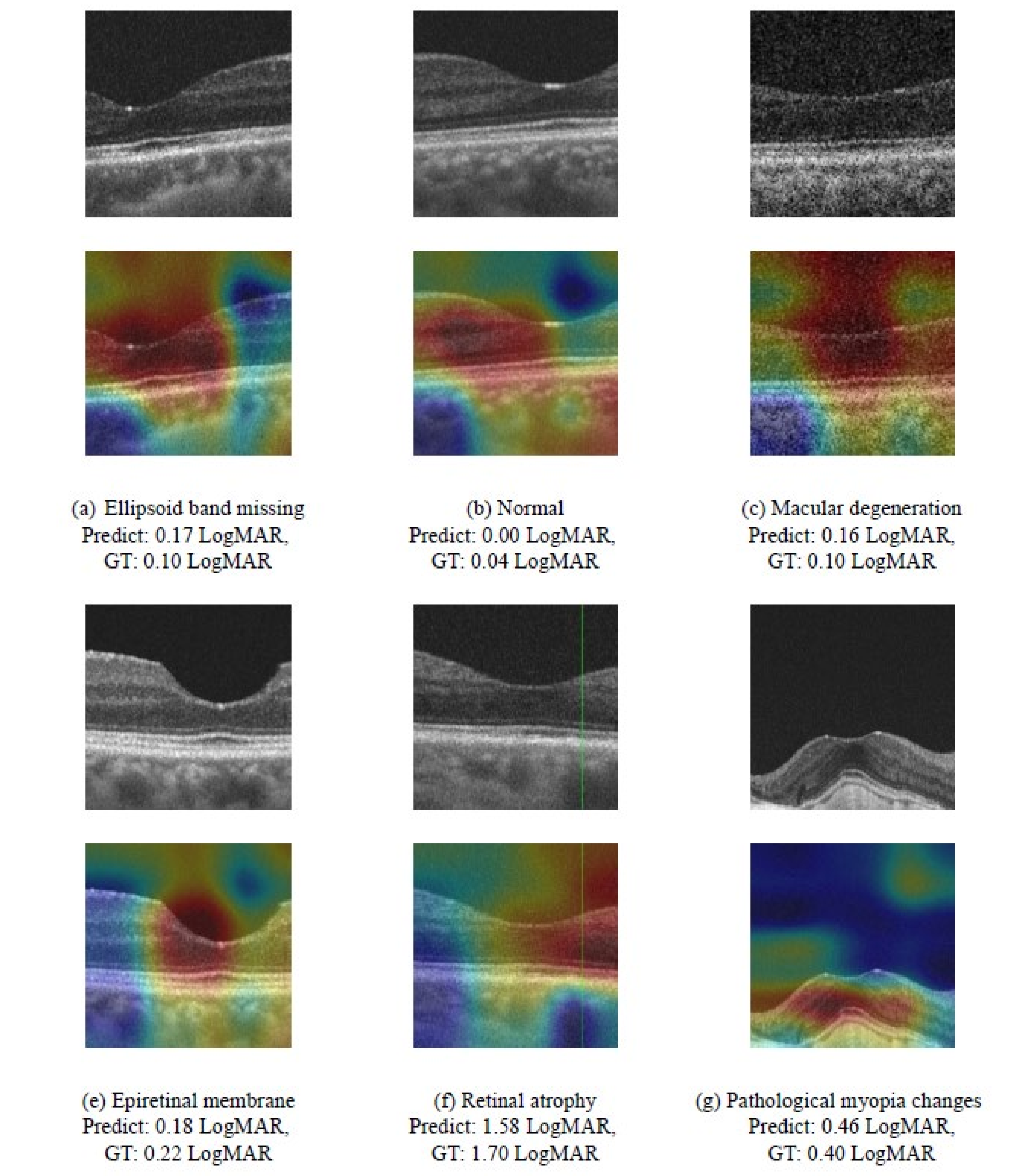


Figure 4. Distribution of predictions and labels.



(a) Ellipsoid band missing
Predict: 0.17 LogMAR,
GT: 0.10 LogMAR

(b) Normal
Predict: 0.00 LogMAR,
GT: 0.04 LogMAR

(c) Macular degeneration
Predict: 0.16 LogMAR,
GT: 0.10 LogMAR

(e) Epiretinal membrane
Predict: 0.18 LogMAR,
GT: 0.22 LogMAR

(f) Retinal atrophy
Predict: 1.58 LogMAR,
GT: 1.70 LogMAR

(g) Pathological myopia changes
Predict: 0.46 LogMAR,
GT: 0.40 LogMAR

Figure 5. Class activation maps (CAM) of different OCT samples.

4. CONCLUSION

Our novel framework capitalizes on multimodal data, bolstering BCVA prediction. The synergistic combination of textual and image data enhances predictive accuracy. Our approach simplifies the handling of incomplete multi-modal datasets, with potential applications beyond our domain.

5. Acknowledgements

This work is partially supported by Bingtuan Science and Technology Program (No. 2022DB005 and 2019BC008) and Key Research and Development Program of Hubei Province (2022BCA009).