
Multimodal Masked Image Modeling for Retinal Image Analysis: An Empirical Study

Qian Wan¹ José Morano^{1,2} Hrvoje Bogunović^{1,2}

¹Institute of Artificial Intelligence, Medical University of Vienna

²Christian Doppler Lab for Artificial Intelligence in Retina, Medical University of Vienna

Abstract

Masked image modeling is a widely used self-supervised learning technique for pretraining and foundation models. In ophthalmology, most existing approaches are either unimodal or unpaired multimodal. In this study, paired color fundus photography (CFP) and optical coherence tomography (OCT) are utilized to investigate the impact of multimodal and unimodal pretraining on the downstream tasks. Experiment results show that multimodal pretraining generally improves downstream performance compared with unimodal pretraining, with notable gains observed on several CFP and OCT datasets.

1 Introduction

Self-supervised learning (SSL) has emerged as a powerful paradigm for medical image analysis. By learning from unlabeled data, SSL reduces models’ reliance on costly expert-level annotations and enables large-scale pretraining. Large-scale pretraining is a key component in developing foundation models, aiming to provide generalizable and transferable features across multiple datasets and tasks.

Masked image modeling (MIM), in which models learn to reconstruct masked regions of an image from partially observed input, has become a dominant SSL paradigm in ophthalmology. Current MIM-based studies follow two paradigms. The first is unimodal pretraining, where models are trained on a single modality such as color fundus photography (CFP) or optical coherence tomography (OCT) [1–4]. The other involves multimodal pretraining with unpaired modalities, where images from different modalities are incorporated either sequentially [5] or through modality mixing [6].

However, CFP and OCT are complementary in capturing surface-level appearance and deeper structural alterations and are often acquired from the same patient during a single examination, resulting in naturally paired multimodal data. Such paired data provides an opportunity to exploit cross-modal correspondence and potentially learn richer model representations. Despite this, the impact of multimodal MIM using paired CFP and OCT remains insufficiently explored.

In this work, we conduct an empirical study to investigate the effect of multimodal MIM with paired CFP and OCT images. Under controlled experimental settings with identical training data, we compare unimodal and multimodal pretraining strategies and evaluate their performance across downstream classification tasks.

2 Method

Following the MIRAGE [7], we use the same architecture as MultiMAE [8], as it has been validated on retinal imaging with paired scanning laser ophthalmoscopy and OCT modalities. Overall, it consists of modality-specific tokenizers, a modality-shared Transformer encoder, and modality-specific Transformer decoders, as shown in Fig. 1. Specifically, each input is mapped to a set of discrete tokens by modality-specific tokenizers. Those tokens are randomly selected as visible and

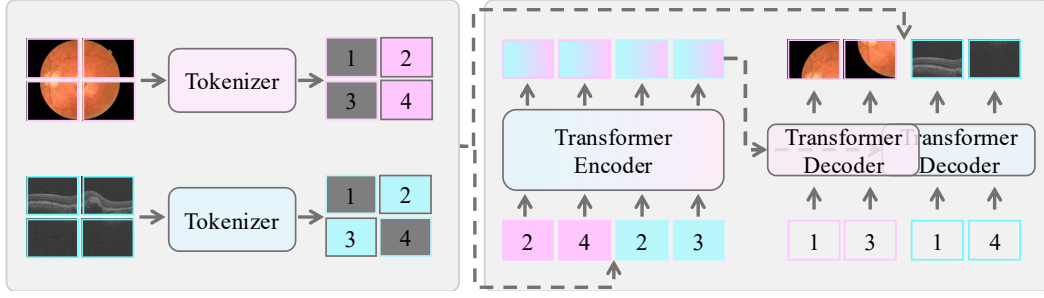


Figure 1: Model architecture. Images are first tokenized, after which a subset of tokens is randomly selected as visible tokens while the remaining tokens are masked (gray-filled). The visible tokens are then processed by a modality-shared Transformer encoder. Finally, placeholder tokens (unfilled) together with the encoded tokens are fed into modality-specific Transformer decoders.

masked tokens. Visible tokens from all modalities are concatenated and fed into the modality-shared Transformer encoder, along with a global token. The encoded visible tokens are then fed into modality-specific Transformer decoders, each with its set of placeholder tokens. The output of each Transformer decoder is optimized to reconstruct the original input corresponding to its mask tokens.

3 Experiments

Datasets. Pretraining uses 11,527 in-house paired CFP and OCT (central slice) images from 1,239 patients, including 3,814 AMD, 2,370 RVO, and 782 DME samples. Evaluation uses seven CFP datasets: PAPILA [9], JSIEC [10], MESSIDOR2 [11], GAMMA [12], Glaucoma Fundus [13], IDRiD [14], and APTOS2019 [15] (Tab. 1), and seven OCT datasets: Duke iAMD [16], Noor Eye Hospital [17], OCTDL [18], OCTID [19], Harvard Glaucoma [20], OLIVES [21], and GAMMA (Tab. 2).

Table 1: Details of the downstream CFP classification datasets.

Dataset	# Samples	# Control	Classes (# Samples)	Disease Type
PAPILA	488	333	Suspect (87), Glaucoma (68)	Glaucoma
JSIEC	1,000	38	38 classes (962)	Mixed
MESSIDOR2	1,744	1,017	Mild (270), Moderate (347) Severe (75), Proliferative (35)	DR
GAMMA	200	101	Early (51), Intermediate/Advanced (48)	Glaucoma
Glaucoma Fundus	1,544	788	Early (289), Advanced (467)	Glaucoma
IDRiD	516	168	Mild (25), Moderate (168) Severe (93), Proliferative (62)	DR
APTOS 2019	3,662	1,805	Mild (370), Moderate (999) Severe (193), Proliferative (295)	DR
Retina	601	300	Cataract (100), Glaucoma (101) Retinal Diseases (100)	Mixed

Implementation details. Multimodal and unimodal pretraining share the same architecture, masking scheme, and training setup, while unimodal pretraining is performed using either CFP-only or OCT-only inputs. The input resolution is set to 512×512 with a patch size of 32×32 . The masking ratio is fixed at 85%. We use the AdamW optimizer with a learning rate of 2.5×10^{-4} . Training runs for 1600 epochs, including a 40-epoch warm-up starting from an initial learning rate of 1×10^{-6} , followed by cosine decay. Pretraining is conducted on 8 H100 GPUs with a batch size of 16 per GPU. Horizontal flipping is applied during training with a probability of 0.5.

Evaluation metrics. Performance is evaluated using five metrics: area under the receiver operating characteristic curve (AUROC), average precision (AP), balanced accuracy (BAcc), F1-score, and

Table 2: Details of the downstream OCT classification datasets.

Dataset	# Samples	# Control	Classes (# Samples)	Disease Type
Duke iAMD	383	115	iAMD (268)	AMD
Noor Eye Hospital	148	50	AMD (48), DME (50)	Mixed
OCTDL	2,063	332	AMD (1,230), ERM (155), DME (147) RVO (101), VID (76), RAO (22)	Mixed
OCTID	572	206	AMD (55), DR (107), CSR (102), MH (102)	Mixed
Harvard Glaucoma	1,000	557	Glaucoma (443)	Glaucoma
OLIVES	1,590	0	DME (931), DR (659)	DR
GAMMA	100	50	Early (26), Intermediate/Advanced (24)	Glaucoma

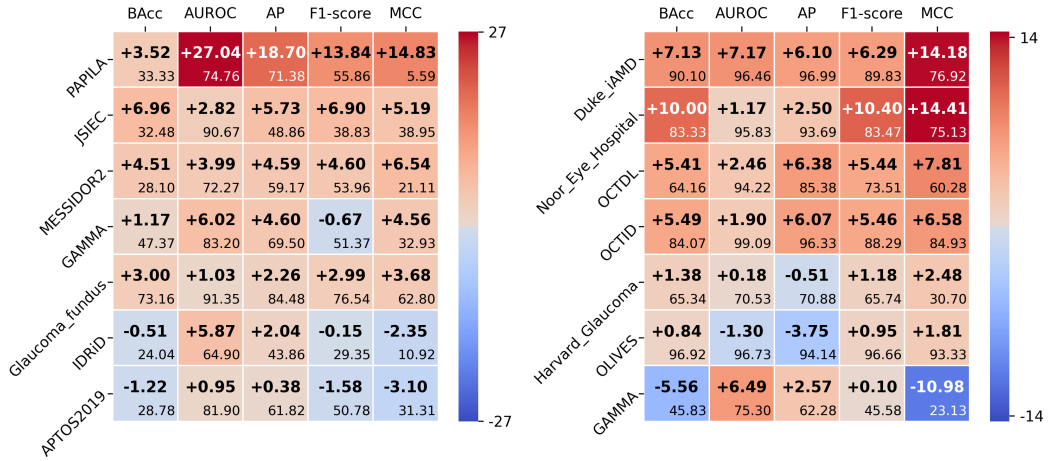


Figure 2: Performance difference of multimodal over unimodal pretraining on downstream classification tasks, evaluated via linear probing. Results are shown for CFP-based classification tasks (left) and OCT-based classification tasks (right). In each cell, the larger bolded value denotes the performance difference (multimodal minus unimodal), while the smaller value denotes the absolute performance of multimodal pretraining. Metrics range from 0 to 100, except MCC (-100 to 100).

Matthews correlation coefficient (MCC). AUROC and AP measure ranking quality, while BAcc, F1-score, and MCC are computed from predictions obtained via argmax over the softmax outputs.

Results. Fig. 2 illustrates the performance improvement of multimodal pretraining compared with unimodal pretraining across multiple datasets and evaluation metrics. Fig. 2(a) shows results for CFP-based classification tasks, while Fig. 2(b) shows results for OCT-based classification tasks. Each row corresponds to a dataset and each column to an evaluation metric. Positive values (red) indicate performance gains from multimodal pretraining, whereas negative values (blue) indicate decreases.

Overall, multimodal pretraining provides consistent and sometimes substantial performance gains, though the magnitude varies across datasets and metrics.

For CFP-based tasks, the largest improvements occur on PAPILA, particularly for AUROC and AP, with consistent gains across all metrics. Moderate improvements are also observed for JSIEC and MESSIDOR2, whereas smaller gains appear for GAMMA and Glaucoma fundus. In contrast, limited or negative gains are observed for IDRiD and APTOS2019. The negative MCC observed in the unimodal setting on PAPILA suggests systematic misclassification, likely due to domain mismatch between optic disc-centered PAPILA images and the macular-centered pretraining data, which linear probing cannot adequately address.

For OCT-based tasks, multimodal pretraining generally improves performance on several datasets. Notable gains are observed on Duke iAMD and Noor Eye Hospital, especially in MCC and F1-score.

Moderate improvements appear on OCTDL and OCTID, while smaller gains or slight decreases are occur on Harvard Glaucoma and OLIVES. The GAMMA dataset shows mixed results with improvements in AUROC and AP but decreases in MCC and BAcc.

4 Discussion

Although multimodal pretraining generally improves downstream performance, the gains are not uniform across datasets. For CFP-based tasks, the decreases on IDRiD and APTOS2019 may be attributed to a data distribution shift, as the pretraining data use the full field of view (FOV) without any preprocessing, whereas images in IDRiD and APTOS2019 are FOV-cropped.

A similar phenomenon appears in OCT-based tasks. While most datasets benefit from multimodal pretraining, the GAMMA dataset shows mixed results, with improvements in ranking-based metrics such as AUROC and AP but decreases in label-based metrics including MCC and BAcc. This pattern suggests that multimodal pretraining improves the separability of positive and negative scores but may alter the score distribution across classes, which can affect the final predicted labels.

Although the pretraining data contain 6,966 samples associated with diseases such as AMD, RVO, and DME, the largest downstream improvements are not necessarily observed on datasets involving these diseases. One explanation is that the pretraining primarily learns generalizable retinal representations rather than disease-specific features. These representations capture structural and pathological patterns shared across retinal conditions, such as texture irregularities, lesion boundaries, or anatomical distortions. Consequently, the learned features can transfer effectively to downstream tasks even when the target diseases differ from those dominating the pretraining data. In addition, differences in dataset characteristics, such as imaging devices, data distributions, annotation protocols, and task difficulty, may influence the magnitude of performance gains across datasets.

Limitations and future work. The pretraining dataset remains relatively limited, containing only 1,239 patients. In addition, a substantial portion of the samples (4,561) lacks class labels. These unlabeled samples may contain cases from various disease categories. As a result, improvements observed in certain downstream tasks may partly stem from similarities between the underlying patterns in the pretraining data and the downstream datasets. Furthermore, the current pretraining dataset provides limited metadata, such as imaging devices or acquisition settings, which may contain additional information useful for understanding such distributional effects and discovering patterns that influence downstream performance. Another limitation is that the downstream evaluation in this study focuses primarily on classification tasks. The effectiveness of multimodal pretraining for other tasks remains unclear.

In this work, multimodal pretraining is conducted using CFP images and central-slice OCT scans. Future work could explore incorporating CFP with full OCT volumes, which may provide richer structural information and further improve multimodal pretraining and downstream task performance. Further investigation of fine-tuning strategies and downstream adaptation protocols may help better exploit pretrained representations.

5 Conclusion

This study investigates multimodal pretraining using paired CFP and OCT images for retinal disease analysis. Experiments across multiple CFP and OCT classification tasks demonstrate that multimodal pretraining generally improves downstream performance compared with unimodal pretraining. The magnitude of improvement varies across datasets, highlighting the influence of data distribution.

Acknowledgments and Disclosure of Funding

Funded by the European Union (ERC, HealthAEye, 101171183). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work was also supported in part by the Christian Doppler Research Association, Austrian Federal Ministry of Economy, Energy and Tourism, and the National Foundation for Research, Technology.

References

- [1] Yeonkyung Lee, Woojung Han, Youngjun Jun, Hyeonmin Kim, Jungkyung Cho, and Seong Jae Hwang. Preti: Patient-aware retinal foundation model via metadata-guided representation learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 523–533. Springer, 2025.
- [2] Zixuan Liu, Hanwen Xu, Addie Woicik, Linda G Shapiro, Marian Blazes, Yue Wu, Verena Steffen, Catherine Cukras, Cecilia S Lee, Miao Zhang, et al. Octcube-m: A 3d multimodal optical coherence tomography foundation model for retinal and systemic diseases with cross-cohort and cross-device validation. *arXiv preprint arXiv:2408.11227*, 2024.
- [3] Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023.
- [4] Jianing Qiu, Jian Wu, Hao Wei, Peilun Shi, Mingqing Zhang, Yunyun Sun, Lin Li, Hanruo Liu, Hongyi Liu, Simeng Hou, et al. Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence. *NEJM AI*, 1(12):A10a2300221, 2024.
- [5] Zhiyuan Cai, Li Lin, Huaqing He, and Xiaoying Tang. Uni4eye: Unified 2d and 3d self-supervised pre-training via masked image modeling transformer for ophthalmic image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 88–98. Springer, 2022.
- [6] Danli Shi, Weiyi Zhang, Xiaolan Chen, Yexin Liu, Jiancheng Yang, Siyu Huang, Yih Chung Tham, Yingfeng Zheng, and Mingguang He. Eyefound: a multimodal generalist foundation model for ophthalmic imaging. *arXiv preprint arXiv:2405.11338*, 2024.
- [7] José Morano, Botond Fazekas, Emese Sükei, Ronald Fecso, Taha Emre, Markus Gumpinger, Georg Faustmann, Marzieh Oghbaie, Ursula Schmidt-Erfurth, and Hrvoje Bogunović. Multimodal foundation model and benchmark for comprehensive retinal oct image analysis. *NPJ Digital Medicine*, 8(1):576, 2025.
- [8] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European conference on computer vision*, pages 348–367. Springer, 2022.
- [9] Oleksandr Kovalyk, Juan Morales-Sánchez, Rafael Verdú-Monedero, Inmaculada Sellés-Navarro, Ana Palazón-Cabanes, and José-Luis Sancho-Gómez. Papila: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. *Scientific Data*, 9(1):291, 2022.
- [10] Ling-Ping Cen, Jie Ji, Jian-Wei Lin, Si-Tong Ju, Hong-Jie Lin, Tai-Ping Li, Yun Wang, Jian-Feng Yang, Yu-Fen Liu, Shaoying Tan, et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nature communications*, 12(1):4828, 2021.
- [11] Etienne Decenci re, Xiwei Zhang, Guy Cazuguel, Bruno Lay, B atrice Cochener, Caroline Trone, Philippe Gain, John-Richard Ord  nez-Varela, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, pages 231–234, 2014.
- [12] Junde Wu, Huihui Fang, Fei Li, Huazhu Fu, Fengbin Lin, Jiongcheng Li, Yue Huang, Qinji Yu, Sifan Song, Xinxing Xu, et al. Gamma challenge: glaucoma grading from multi-modality images. *Medical Image Analysis*, 90:102938, 2023.
- [13] Jin Mo Ahn, Sangsoo Kim, Kwang-Sung Ahn, Sung-Hoon Cho, Kwan Bok Lee, and Ungsoo Samuel Kim. A deep learning model for the detection of both advanced and early glaucoma using fundus photography. *PLoS one*, 13(11):e0207982, 2018.
- [14] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabudhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018.
- [15] Karthik, Maggie, and Sohier Dane. Aptos 2019 blindness detection. <https://kaggle.com/competitions/aptos2019-blindness-detection>, 2019. Kaggle.
- [16] Sina Farsiu, Stephanie J Chiu, Rachelle V O’Connell, Francisco A Folgar, Eric Yuan, Joseph A Izatt, Cynthia A Toth, Age-Related Eye Disease Study 2 Ancillary Spectral Domain Optical Coherence Tomography Study Group, et al. Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography. *Ophthalmology*, 121(1):162–172, 2014.

- [17] Reza Rasti, Hossein Rabbani, Alireza Mehrdehnavi, and Fedra Hajizadeh. Macular oct classification using a multi-scale convolutional neural network ensemble. *IEEE transactions on medical imaging*, 37(4): 1024–1034, 2017.
- [18] Mikhail Kulyabin, Aleksei Zhdanov, Anastasia Nikiforova, Andrey Stepichev, Anna Kuznetsova, Mikhail Ronkin, Vasilii Borisov, Alexander Bogachev, Sergey Korotkich, Paul A Constable, et al. Octdl: Optical coherence tomography dataset for image-based deep learning methods. *Scientific data*, 11(1):365, 2024.
- [19] Peyman Gholami, Priyanka Roy, Mohana Kuppuswamy Parthasarathy, and Vasudevan Lakshminarayanan. Octid: Optical coherence tomography image database. *Computers & Electrical Engineering*, 81:106532, 2020.
- [20] Yan Luo, Min Shi, Yu Tian, Tobias Elze, and Mengyu Wang. Harvard glaucoma detection and progression: A multimodal multitask dataset and generalization-reinforced semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20471–20482, 2023.
- [21] Mohit Prabhushankar, Kiran Kokilepersaud, Yash-ye Logan, Stephanie Trejo Corona, Ghassan AlRegib, and Charles Wykoff. Olives dataset: Ophthalmic labels for investigating visual eye semantics. *Advances in Neural Information Processing Systems*, 35:9201–9216, 2022.