1 Supplementary Material

1.1 Multimodal Pre-training Methods

Medical vision-language pre-training enhances medical image analysis by learning domain-specific features from medical images paired with clinical descriptions. By jointly encoding images and reports, these models better understand visual and textual information, improving performance and interpretability. Typical methods improve image-text contrastive learning [3,7,18,34], align image and text embeddings using semantic labels [31], or enhance image representation through masked image and language modeling [36]. Recent methods have focused on radiology, especially chest X-rays [22,32,33], due to the abundance of image-report pairs that help learn the relationship between visual features and medical findings. However, this approach is less applicable in other medical domains like ophthalmology, where retinal images have diverse modalities and generally lack accompanying text information.

Unlike RETFound [37] and FLAIR [27], we propose a universal retinal FM that processes multiple imaging modalities and integrates various expert annotations into the image encoder. By leveraging multimodal images and domain knowledge, this model enables comprehensive representations, facilitates multimodal reasoning, captures broader anatomical and physiological relationships, and reduces development and maintenance costs.

1.2 Dataset Preparation

Pre-training Dataset. Based on FLAIR, we collected a large dataset (Tabel. 1) comprising 187,270 publicly accessible CFP and OCT images for the pre-training of our foundation model and the experiments conducted. More details can be found in FLAIR [27].

Fine-tuning Dataset. To conduct a comprehensive evaluation of the foundation model, we collected 7 CFP datasets and 1 OCT dataset according to the experimental setup defined by RETFound, and divided them following the data division ratios provided by RETFound [37].

Task Specific Dataset. Based on the labels in the pre-training dataset, we constructed a task-specific dataset for Diabetic Retinopathy classification, which includes images from EYEPACS, PARAGUAY, OIA-DDR, and DeepDRiD, totaling 51,556 images. Similarly, a task-specific dataset for OCT disease classification was developed based on the OCTCELL dataset.

1.3 Expert Knowledge Descriptions

For the domain knowledge descriptors related to retinal diseases based on CFP, we referred to FLAIR [27] for guidance. Meanwhile, for the domain knowledge descriptors concerning retinal diseases based on OCT, we utilized ChatGPT-4 to summarize four distinct descriptions for the corresponding disease label names, which were then employed as the domain knowledge descriptors (Tabel. 2).

Table 1: Collected publicly available dataset for foundation model pre-training.

No.	Name	Count	Labels
1	OCTCELL [16]	83,484	CNV, DME, DRUSEN, and NORMAL
2	EYEPACS[11]	35,126	noDR, mildDR, modDR, sevDR, prolDR
3	RFMid [25]	3,170	DR, ARMD, MH, DN, MYA, BRVO, TSLN, ERM, LS, MS CSR, ODC, CRVO, TV, AH, ODP, ODE, ST, AION, PT, RT RS, CRS, EX, RPEC, MHL, RP, CWS, CB, ODM, PRH, MNF, HR, CRAO, TD, CME, PTCR, CF, VH, MCA VS, BRAO, PLQ, HPED, CL
4	EYENET [15]	15,709	*Text*
5	LAG [19]	4,854	G, noG
6	ODIR [1]	10,846	N, DR, G, CAT, ARMD, HR, MYA
7	PARAGUAY [4]	757	noDR, mildDR, modDR, sevDR, prolDR
8	STARE [14]	397	*Text*
9	ARIA [12]	143	N, ARMD, DR
10	AGAR300 [9]	28	DR, MA
11	FUND-OCT [13]	179	G, N, CME, neovARMD, geoARMD, acCSR, chCSR
12	DRIONS-DB [6]	110	noCAT, Dis
13	Drishti-GS1 [28]	101	N, G
14	E-ophta [8]	265	EX, MA
15	G1020 [2]	1,020	G, N
16	HRF [5]	45	N, G, DR, noisy
17	ORIGA [35]	650	G, noG
18	ROC [24]	100	MA
19	OIA-DDR [20]	13,673	noDR, mildDR, modDR, sevDR, prolDR, HE, hEX, sEX, MA
20	SYSU [21]	1,219	noDR, mildDR, modDR, sevDR, prolDR, HE, hEX, sEX
21	JICHI [29]	9,939	noDR, mildDR, modDR, sevDR, prolDR
22	CHAKSU [17]	284	G, noG
23	DR1-2 [26]	1,469	N, ReSD, hEX, DN, CWS, supHE, deepHE
24	ScarDat [30]	997	LS, noLS
25	ACRIMA [10]	705	G, noG
26	DeepDRiD [23]	2,000	noDR, mildDR, modDR, sevDR, prolDR
	Total	187,270	

1.4 Statistical Significance Analysis

Fig. 1 shows the statistically significant analysis of UrFound compared to the second-best results in Table 1 of the paper, based on a t-test with a p-value of 0.05. UrFound performs similarly to the second-best method on IDRID and JSIEC, and significantly better on the other six datasets.

1.5 External Validation

We conducted external evaluations on the IDRID, APTOS, and Messidor datasets and found that our UrFound model demonstrates strong generalizability, and outperforms RETFound and FLAIR in most cases, with statistical significance based on a t-test with a p-value of 0.05 (Fig. 2).

References

1. Peking University - ODIR 2019: Ophthalmic Disease Intelligent Recognition. `https://odir2019.grand-challenge.org/background/`, accessed: 2024-01-04
2. Bajwa, M.N., Singh, G.A.P., Neumeier, W., Malik, M.I., Dengel, A., Ahmed, S.: G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma

Table 2: Expert Knowledge descriptions for OCT-based retinal diseases.

Category	Domain Knowledge descriptor
CNV	1. “The OCT image reveals a network of new blood vessels beneath the retinal pigment epithelium, indicative of Choroidal Neovascularization. These vessels are irregular and often associated with age-related macular degeneration." 2. “There is noticeable distortion and elevation of the overlying retinal layers, which is characteristic of the leakage and bleeding from these abnormal vessels." 3. “Pockets of fluid accumulation under the retina, known as subretinal fluid, are evident, causing a dome-shaped elevation of the retina." 4. “Areas of hemorrhage and exudation are visible between the retinal layers and beneath the retinal pigment epithelium, indicating active vascular leakage."
DME	1. “In Diabetic Macular Edema, the OCT scan shows a significant thickening of the macula, particularly in the inner retinal layers, due to fluid accumulation. This condition is a common complication of diabetic retinopathy." 2. “Multiple cystic spaces within the retinal layers are observed, filled with fluid, giving a sponge-like appearance to the retina." 3. “Hyperreflective foci are seen below the retinal pigment epithelium, representing hard exudates, which are residues of lipid deposits from leaking blood vessels." 4. “In advanced cases, disruption and irregularity of the retinal pigment epithelium layer are noted, likely due to chronic edema and vascular leakage."
DRUSEN	1. “Drusen appear as small, round elevations beneath the retinal pigment epithelium layer in OCT images. These are accumulations of extracellular material, commonly associated with age-related macular degeneration." 2. “The drusen vary in size and confluence, with larger and more numerous drusen indicating a higher risk of progression to advanced macular degeneration." 3. “In cases of extensive drusen, there is noticeable distortion and thickening of the overlying retinal pigment epithelium layer." 4. “Some drusen exhibit a central hyperreflective core with a surrounding hyporeflective halo, suggesting varying stages of drusen evolution."
NORMAL	1. “The normal retina in OCT imaging presents a well-defined, multi-layered structure. Each layer exhibits its characteristic reflectivity, with clear demarcation between layers." 2. “The retinal pigment epithelium layer appears as a uniform, thin band adjacent to the highly reflective Bruch’s membrane." 3. “The photoreceptor layer, including the cones and rods, is orderly and shows no signs of fluid accumulation or structural distortion." 4. “The nerve fiber layer, ganglion cell layer, and inner and outer nuclear layers all display normal thickness and reflectivity, with no signs of pathology or abnormality."

detection. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–7. IEEE (2020)
3. Bazi, Y., Rahhal, M.M.A., Bashmal, L., Zuair, M.: Vision–language model for visual question answering in medical imagery. Bioengineering **10**(3), 380 (2023)
4. Benítez, V.E.C., Matto, I.C., Román, J.C.M., Noguera, J.L.V., García-Torres, M., Ayala, J., Pinto-Roa, D.P., Gardel-Sotomayor, P.E., Facon, J., Grillo, S.A.: Dataset from fundus images for the study of diabetic retinopathy. Data in brief **36**, 107068 (2021)
5. Budai, A., Bock, R., Maier, A., Hornegger, J., Michelson, G., et al.: Robust vessel segmentation in fundus images. International journal of biomedical imaging **2013** (2013)
6. Carmona, E.J., Rincón, M., García-Feijoó, J., Martínez-de-la Casa, J.M.: Identification of the optic nerve head with genetic algorithms. Artificial intelligence in medicine **43**(3), 243–259 (2008)
7. Chen, Z., Diao, S., Wang, B., Li, G., Wan, X.: Towards unifying medical vision-and-language pre-training via soft prompts. arXiv preprint arXiv:2302.08958 (2023)

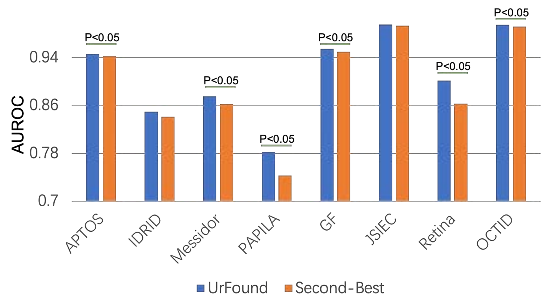

Fig. 1: Analysis of Statistical Significance with the Second-Best Results in Table 1 of the Paper.

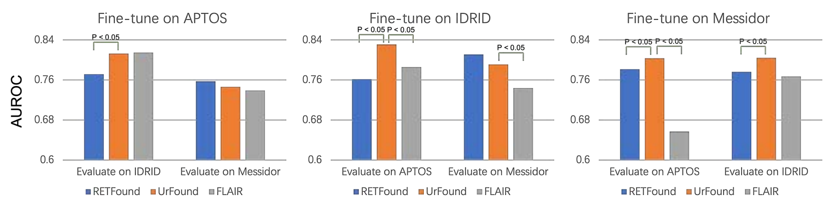

Fig. 2: Performance of RETFound, UrFound, and FLAIR on External Validation with Statistical Analysis

8. Decenciere, E., Cazuguel, G., Zhang, X., Thibault, G., Klein, J.C., Meyer, F., Marcotegui, B., Quellec, G., Lamard, M., Danno, R., et al.: Teleophta: Machine learning and image processing methods for teleophthalmology. Irbm **34**(2), 196–203 (2013)
9. Derwin, D.J., Selvi, S.T., Singh, O.J., Shan, B.P.: A novel automated system of discriminating microaneurysms in fundus images. Biomedical Signal Processing and Control **58**, 101839 (2020)
10. Diaz-Pinto, A., Morales, S., Naranjo, V., Köhler, T., Mossi, J.M., Navea, A.: Cnns for automatic glaucoma assessment using fundus images: an extensive validation. Biomedical engineering online **18**, 1–19 (2019)
11. Emma Dugas, Jared, J.W.C.: Diabetic retinopathy detection (2015), `https://kaggle.com/competitions/diabetic-retinopathy-detection`
12. Farnell, D.J., Hatfield, F.N., Knox, P., Reakes, M., Spencer, S., Parry, D., Harding, S.P.: Enhancement of blood vessels in digital fundus photographs via the application of multiscale line operators. Journal of the Franklin institute **345**(7), 748–765 (2008)
13. Hassan, T., Akram, M.U., Werghi, N., Nazir, M.N.: Rag-fw: A hybrid convolutional framework for the automated extraction of retinal lesions and lesion-influenced grading of human retinal pathology. IEEE journal of biomedical and health informatics **25**(1), 108–120 (2020)
14. Hoover, A., Kouznetsova, V., Goldbaum, M.: Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Transactions on Medical imaging **19**(3), 203–210 (2000)

15. Huang, J.H., Yang, C.H.H., Liu, F., Tian, M., Liu, Y.C., Wu, T.W., Lin, I., Wang, K., Morikawa, H., Chang, H., et al.: Deepopht: medical report generation for retinal images via deep models and visual explanation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2442–2452 (2021)
16. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. cell **172**(5), 1122–1131 (2018)
17. Kumar, J.H., Seelamantula, C.S., Gagan, J., Kamath, Y.S., Kuzhuppilly, N.I., Vivekanand, U., Gupta, P., Patil, S.: Chákṣu: A glaucoma specific fundus image database. Scientific data **10**(1), 70 (2023)
18. 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)
19. Li, L., Xu, M., Wang, X., Jiang, L., Liu, H.: Attention based glaucoma detection: A large-scale database and cnn model. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10571–10580 (2019)
20. Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., Kang, H.: Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Information Sciences **501**, 511–522 (2019)
21. Lin, L., Li, M., Huang, Y., Cheng, P., Xia, H., Wang, K., Yuan, J., Tang, X.: The sustech-sysu dataset for automated exudate detection and diabetic retinopathy grading. Scientific Data **7**(1), 409 (2020)
22. Liu, C., Shah, A., Bai, W., Arcucci, R.: Utilizing synthetic data for medical vision-language pre-training: Bypassing the need for real images. arXiv preprint arXiv:2310.07027 (2023)
23. Liu, R., Wang, X., Wu, Q., Dai, L., Fang, X., Yan, T., Son, J., Tang, S., Li, J., Gao, Z., et al.: Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. Patterns **3**(6) (2022)
24. Niemeijer, M., Van Ginneken, B., Cree, M.J., Mizutani, A., Quellec, G., Sánchez, C.I., Zhang, B., Hornero, R., Lamard, M., Muramatsu, C., et al.: Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. IEEE transactions on medical imaging **29**(1), 185–195 (2009)
25. Pachade, S., Porwal, P., Thulkar, D., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Giancardo, L., Quellec, G., Mériaudeau, F.: Retinal fundus multi-disease image dataset (rfmid): A dataset for multi-disease detection research. Data **6**(2), 14 (2021)
26. Pires, R., Jelinek, H.F., Wainer, J., Valle, E., Rocha, A.: Advancing bag-of-visual-words representations for lesion classification in retinal images. PloS one **9**(6), e96814 (2014)
27. Silva-Rodriguez, J., Chakor, H., Kobbi, R., Dolz, J., Ayed, I.B.: A foundation language-image model of the retina (FLAIR): Encoding expert knowledge in text supervision. arXiv preprint arXiv:2308.07898 (2023)
28. Sivaswamy, J., Krishnadas, S., Joshi, G.D., Jain, M., Tabish, A.U.S.: Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In: 2014 IEEE 11th international symposium on biomedical imaging (ISBI). pp. 53–56. IEEE (2014)
29. Takahashi, H., Tampo, H., Arai, Y., Inoue, Y., Kawashima, H.: Applying artificial intelligence to disease staging: Deep learning for improved staging of diabetic retinopathy. PloS one **12**(6), e0179790 (2017)

30. Wei, Q., Li, X., Wang, H., Ding, D., Yu, W., Chen, Y.: Laser scar detection in fundus images using convolutional neural networks. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14. pp. 191–206. Springer (2019)
31. Yan, B., Pei, M.: Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2982–2990 (2022)
32. You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: Cxr-clip: Toward large scale chest x-ray language-image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023)
33. Zhang, X., Wu, C., Zhang, Y., Xie, W., Wang, Y.: Knowledge-enhanced visual-language pre-training on chest radiology images. Nature Communications **14**(1), 4542 (2023)
34. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare Conference. pp. 2–25. PMLR (2022)
35. Zhang, Z., Yin, F.S., Liu, J., Wong, W.K., Tan, N.M., Lee, B.H., Cheng, J., Wong, T.Y.: Origa-light: An online retinal fundus image database for glaucoma analysis and research. In: 2010 Annual international conference of the IEEE engineering in medicine and biology. pp. 3065–3068. IEEE (2010)
36. Zhou, H.Y., Lian, C., Wang, L., Yu, Y.: Advancing radiograph representation learning with masked record modeling. In: Proceddings of ICLR. pp. 1–16 (2023)
37. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., et al.: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023)