

MM-Retinal: Knowledge-Enhanced Foundational Pretraining with Fundus Image-Text Expertise

Ruiqi Wu^{1,2}, Chenran Zhang^{1,2}, Jianle Zhang^{1,2}, Yi Zhou^{1,2*},
Tao Zhou³, and Huazhu Fu⁴

¹ School of Computer Science and Engineering, Southeast University, China

² Key Laboratory of New Generation Artificial Intelligence Technology and Its

Interdisciplinary Applications, Ministry of Education, China

³ Nanjing University of Science and Technology, Nanjing, China

⁴ Agency for Science, Technology and Research (A*STAR), Singapore

{ruiqiwu@seu.edu.cn, yizhou.szcn@gmail.com}

Abstract. Current fundus image analysis models are predominantly built for specific tasks relying on individual datasets. The learning process is usually based on data-driven paradigm without prior knowledge, resulting in poor transferability and generalizability. To address this issue, we propose **MM-Retinal**, a multi-modal dataset that encompasses high-quality image-text pairs collected from professional fundus diagram books. Moreover, enabled by MM-Retinal, we present a novel **Knowledge-enhanced foundational pretraining model** which incorporates **Fundus Image-Text expertise**, called **KeepFIT**. It is designed with image similarity-guided text revision and mixed training strategy to infuse expert knowledge. Our proposed fundus foundation model achieves state-of-the-art performance across six unseen downstream tasks and holds excellent generalization ability in zero-shot and few-shot scenarios. MM-Retinal and KeepFIT are available at <https://github.com/lxirich/MM-Retinal>.

Keywords: Fundus image · Foundational pretraining · Knowledge-enhanced

1 Introduction

Deep learning has achieved great progress in fundus image analysis. However, most previous works [6, 18, 30, 34] usually utilize individual datasets to train task-specific models. This fashion results in three major model weaknesses: 1) poor generalization ability and robustness across varying scenarios; 2) lack of professional fundus domain-knowledge guidance in learning phase; 3) a huge demand for annotated training data. These challenges underscore the need for developing a general-purpose foundation model which is able to analyze comprehensive ocular diseases in fundus image area. Moreover, learning such a model with less training data but more prior knowledge is preferred.

* Corresponding author: Yi Zhou

In fundus image field, there are many specific image-only public datasets for diverse ocular diseases, such as glaucoma [16, 20], diabetic retinopathy [4, 5, 14, 33], age-related macular degeneration [9], and pathological myopia [8]. Despite remarkable progress of foundation models has been witnessed in many fields, such as natural images [29, 36], medical images like chest X-rays [21, 26] and MRI [15, 19], it lags far behind in fundus image area. Thus, it is of considerable value to explore foundational pretraining in this area. RETFound [35] and FLAIR [25] made two preliminary attempts but still suffer from certain limitations compared to other foundation models. RETFound only relies on large-scale image data and adopts a masked image modeling manner. FLAIR exploits both vision and language modalities through contrastive pretraining objective, yet it simply maps category label names to fixed texts. Both of them still lacks integrating fundus expertise with rich and profound image-text descriptions.

To accomplish our intention, a high-quality vision-language fundus dataset containing expert knowledge is required. Such a dataset aims to not only promote the development of foundational fundus models with strong generalizability, robustness, and transferability, but also advance research in incorporating knowledge into learning more interpretable models without manual label annotation. Additionally, it should propel research of multimodal fundus image analysis and beyond. Therefore, we built **MM-Retinal**, a multi-modal dataset which comprises image-text paired data of color fundus photography (CFP), fundus fluorescein angiography (FFA), and optical coherence tomography (OCT) images. All these data are collected from **fundus diagram books** containing comprehensive ocular knowledge, with accurate image-text descriptions provided by ophthalmologists. Moreover, we also developed **KeepFIT**, a knowledge-enhanced foundation model, enabled by MM-Retinal. Specifically, image similarity-guided text revision method and mixed training strategy are proposed to inject fundus expert knowledge from our constructed MM-Retinal into model training.

We highlight our main contributions: **1)** We construct a multi-modal MM-Retinal with over 4.3K high-quality image-text pairs in CFP, FFA and OCT modalities. These pairs are accurately matched with long texts, extensive vocabulary and comprehensive ocular disease and abnormalities. **2)** A knowledge-enhanced foundation model, KeepFIT, is proposed including a vision-language pre-training framework and knowledge integration methods, termed image similarity-guided text revision and mixed training strategy. **3)** KeepFIT achieves state-of-the-art performance on six representative downstream tasks, especially in zero-shot and few-shot scenarios, demonstrating consistently strong robustness, generalizability and transferability.

2 The MM-Retinal Dataset

2.1 Dataset Construction

To construct MM-retinal with high-quality image-text pairs covering CFP, FFA, and OCT modalities, as shown in Fig. 1(a), our designed semi-automatic pipeline of dataset construction contains four steps. For further details, please refer to the

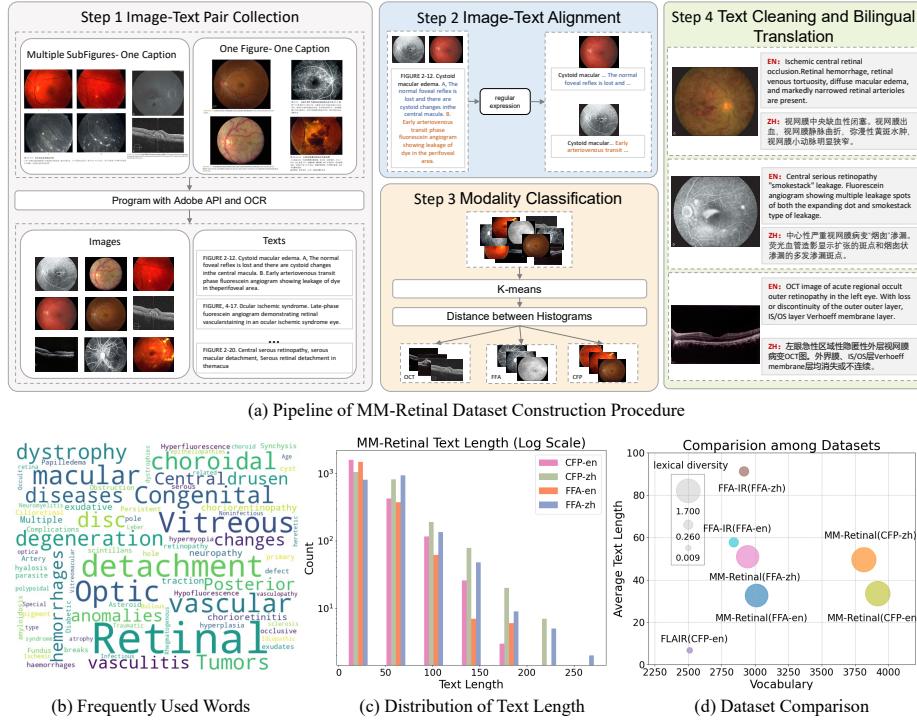


Fig. 1: Construction workflow and statistical overview of MM-Retinal.

supplementary material. A six-person team took four weeks to get MM-Retinal completed.

Step 1: Image-Text Pair Collection: First, image-text pairs are captured from the books. Note that we keep the resolution of each image in the pair no less than 800×800 . Afterwards, Adobe API and OCR techniques are used for raw image and text extraction. **Step 2: Image-Text Alignment:** As mentioned earlier, not all the initially extracted images and texts by the program are very well aligned, as there are some sub-figures correspond to one caption within a screenshot. Thus, we use regular expression matching for sub-figure caption separation to align image-text pairs. **Step 3: Modality Classification:** Since the fundus images in the books include multiple modalities, we categorize images into CFP, FFA, and OCT using K-means and color histogram analysis, and excluding unusual modalities with scarce samples. **Step 4: Text Cleaning and Bilingual Translation:** After manually correcting OCR errors and irrelevant texts, we translate the texts to provide bilingual (English and Chinese) versions, aiming to standardize the language and enhance the influence of MM-Retinal since the diagram books we used include both Chinese and English.

2.2 Dataset Statistics

Current version of MM-Retinal dataset includes 2,169 CFP cases, 1,947 FFA cases and 233 OCT cases. Each case is provided with an image and texts in both English and Chinese. Due to the small scale of OCT modality, we do not explore it for now and will extend this part in future. Detailed statistical analysis of the CFP and FFA data is provided in the Fig. 1, focusing on aspects of frequently used words, text length, vocabulary size, and comparison with other datasets.

Frequently Used Words: Since MM-Retinal dataset is built based on comprehensive ocular fundus diagram books, it covers a wide range of disease categories, such as macular, retinal vascular, and optic nerve diseases. In Fig. 1(b), we only show a small part of the words frequently appeared in CFP and FFA modalities. More retinal abnormal fundus changes and disease categories can be found in the supplementary material.

Text Length: Fig. 1(c) illustrates the text length distribution of our dataset. About 75% of English texts and 45% of Chinese texts range from 1 to 40 words, while 19% of English texts and 43% of Chinese texts contain between 41 and 80 words. Since our dataset is sourced from ophthalmologists' diagram books, it features much longer texts compared to FLAIR which simply maps class names of public datasets into fixed texts.

Vocabulary Size: MM-Retinal dataset contains diverse textual descriptions, such as disease diagnosis, lesion characteristics (e.g. color, shape, appearance), clinical manifestations, and post-treatment efficacy information. Fig. 1(d) showcases the average vocabulary size and lexical diversity across different modalities and languages, where average vocabulary size represents the total number of unique words contained in the whole texts within the dataset and lexical diversity refers to the vocabulary size averaged over each text.

Discussion: Compared to public fundus image datasets, our MM-Retinal stands out from four aspects: **1) Multi-modality:** MM-Retinal is the pioneering fundus dataset that includes high-quality image-text expertise data for CFP, FFA, and OCT. **2) Data Quality.** In contrast to other medical image datasets that often contain low-quality images and captions sourced from websites or academic papers, MM-Retinal provides high-quality images with a resolution over 800×800 , accompanied by accurate and pertinent text descriptions. The images closely match clinical data with minimal domain shift. **3) Category Variety.** According to the contents of four diagram books, MM-Retinal covers a broad range of categories with over 96 abnormalities and diseases. **4) Text Diversity.** MM-Retinal encompasses diverse vocabulary and long texts, containing extensive expert knowledge, and can be explored to enhance data-driven fundus image analysis models.

3 Knowledge-Enhanced Foundational Pretraining

3.1 Vision-Language Pre-training Framework

As shown in Fig. 2, we propose **KeepFIT**, a knowledge-enhanced foundational model, pretrained on public and MM-Retinal datasets. In this paper, we define

public datasets with only category-level labels as unimodal dataset and we follow FLAIR [25] to fill category-level labels into a prompt template to create texts. We utilize ResNet50 [11] as image encoder E_v and BioClinicalBert [3] as text encoder E_t , followed by projection heads P_v and P_t to match image and text feature dimensions d . Thus, given a image x_i and a text y_j , the extracted image features v_i and text features t_j are:

$$v_i = P_v \circ E_v(x_i) \in \mathbb{R}^d, \quad t_j = P_t \circ E_t(y_j) \in \mathbb{R}^d. \quad (1)$$

Given that texts detail the associated fundus images, our goal is to maximize similarity between paired image-text and minimize similarity for unpaired ones in a multimodal space. For MM-retinal, which are genuine texts rather than textual prompts, we follow CLIP [22] to implement the contrastive loss as shown in Eq. 2, and the matching labels G_{v2t}^m and G_{t2v}^m are two $|\mathcal{B}| \times |\mathcal{B}|$ identity matrices:

$$\mathcal{L}_m = \mathbb{E}_{(x,y) \sim \mathcal{B}}[CE(G_{v2t}^m, S_{v2t}) + CE(G_{t2v}^m, S_{t2v})], \quad G_{v2t}^m, G_{t2v}^m \in I_{|\mathcal{B}|}, \quad (2)$$

where m represents our proposed MM-Retinal, \mathcal{B} is the batchsize, CE denotes InfoNCE loss, $S_{v2t} = \lambda(v_i t_j^\top)$ and $S_{t2v} = \lambda(v_i^\top t_j)$ are the cross-modality similarity, λ is a learnable scaling factor, $v2t$ and $t2v$ denote image-to-text and text-to-image, respectively.

For public datasets that only have prompt texts mapped from category-level labels, we follow FLAIR to calculate the category co-occurrence relationships among samples within a batch and construct a target matrix to encourage the pairs belonging to the same category closer. Thus, the objective function Eq. 2 is converted into:

$$\mathcal{L}_p = \mathbb{E}_{(x,y) \sim \mathcal{B}}[CE(G_{v2t}^p, S_{v2t}) + CE(G_{t2v}^p, S_{t2v})], \quad (3)$$

$$G_{v2t}^p = G_{t2v}^p = \begin{cases} 1, & \text{if category}_v = \text{category}_t \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where p represents public datasets that only have category-level labels, the matching labels G_{v2t}^p, G_{t2v}^p are $|\mathcal{B}| \times |\mathcal{B}|$ symmetric matrices.

3.2 Expert Knowledge Integration Methods

As our proposed MM-Retinal features high-quality image-text pairs, long text description, rich vocabulary, and comprehensive ocular diseases categories, it encapsulates extensive fundus expert knowledge. Inspired by TipAdapter [31], we propose a lightweight expert knowledge integration method, called **Image Similarity-Guided Text Revision** along with **Mixed Training Strategy**, to boost the expert knowledge from our MM-Retinal dataset.

Image Similarity-Guided Text Revision: The descriptions within MM-Retinal are more comprehensive and professional than the simple texts mapped

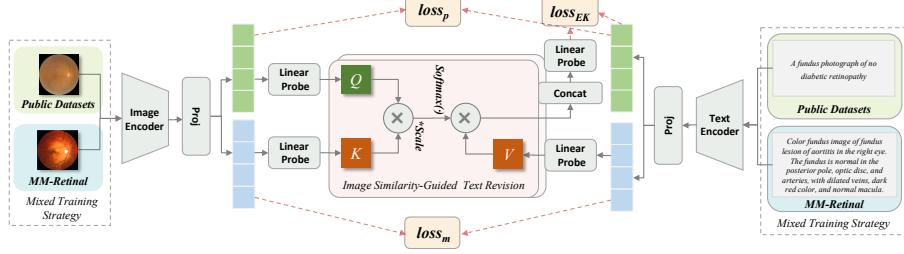


Fig. 2: **KeepFIT**: A vision-language pretraining framework using image-guided text revision and mixed training methods to infuse expert knowledge.

from class names in public datasets. Despite this, the images from both sources share a notable similarity. Hence, we start from identifying visual features in the public datasets that resemble those in MM-Retinal. Visual similarities are used as guidance to extract relevant prior knowledge from MM-Retinal’s text features to refine and enhance the textual prompts of the public datasets.

Specifically, given an input image-text pair $[x_p, y_p]$ from public datasets and $[x_m, y_m]$ from MM-Retinal, the extracted features are (v_p, t_p) and (v_m, t_m) . A multi-head cross-attention [27] is applied to exploit prior expert knowledge as:

$$EK = \text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h)W^O, \quad (5)$$

$$\text{head}_i = \text{ATTN}(v_p W_i^Q, v_m W_i^K, t_m W_i^V), \quad (6)$$

where W_i^Q, W_i^K, W_i^V and W^O are parameter matrices for projection, v_p, v_m, t_m refer to image features of public datasets, image features of MM-Retinal and text features of MM-Retinal, respectively. Then, we establish an expert knowledge revision loss based on Mean Squared Error (MSE) to refine the text features of public datasets by incorporating expert knowledge, as formulated in Eq. 7:

$$\mathcal{L}_{EK} = \text{MSELoss}(EK, t_p). \quad (7)$$

Mixed Training Strategy: Since the public datasets has relatively homogeneous text prompts and almost no expert knowledge, which is vastly different from the texts of our dataset, we propose a mixed training strategy to avoid model optimization bias during the training process. To detail, the samples from public datasets and our dataset are in a 1:1 ratio in each batch.

Overall Training Objective: The overall training objective comprises three parts, which are public datasets contrastive loss \mathcal{L}_p , MM-Retinal contrastive loss \mathcal{L}_m and expert knowledge revision loss \mathcal{L}_{EK} :

$$\mathcal{L} = \mathcal{L}_p + \mathcal{L}_m + \alpha \mathcal{L}_{EK}, \quad (8)$$

where α is to weight for \mathcal{L}_{EK} , empirically set as 100 showing best performance.

Table 1: Comparison of generalization ability on Few-Shot and Zero-Shot tasks. FLAIR and flair denote the model and datasets in [25], respectively. SynFuduns-1M abbreviated as syn. MM represents MM-Retinal.

Method	Data	Few-Shot										Zero-Shot			Avg	
		ODIR200×3										REFUGE FIVES ODIR200×3				
		ClipAdapter			TipAdapter			TipAdapter-f				AUC	ACA	ACA		
		1	5	10	1	5	10	1	5	10						
		ACA			ACA			ACA								
FLAIR	flair	0.720	0.823	0.863	0.403	0.413	0.422	0.417	0.462	0.535	0.926	0.670	0.403	0.588		
	flair+syn	0.735	0.827	0.852	0.603	0.622	0.632	0.580	0.647	0.672	0.880	0.617	0.520	0.682		
KeepFIT	flair	0.763	0.848	0.847	0.780	0.782	0.795	0.775	0.785	0.803	0.931	0.666	0.768	0.795		
	flair+syn	0.795	0.858	0.862	0.751	0.777	0.783	0.760	0.795	0.807	0.856	0.696	0.777	0.793		
	50%flair+MM	0.832	0.873	0.887	0.862	0.870	0.872	0.870	0.883	0.873	0.934	0.654	0.862	0.856		
	flair+MM	0.848	0.878	0.893	0.823	0.843	0.842	0.820	0.847	0.853	0.941	0.731	0.812	0.844		

4 Experiment and Results

4.1 Datasets and Baselines

Pre-training Data: 1) **flair** [25] (**CFP**) compiles 37 open-access fundus image datasets covering 96 categories with up to 284,660 images. These datasets provide category-level labels for classification. 2) **SYNFUNDUS-1M** [24] (**CFP**) is a synthetic dataset with 1 million images for 14 diseases, created by a diffusion model [12] trained on 1.3 million private fundus images. 3) **FFA-IR** [17] (**FFA**) provides 10,790 reports along with 1,048,584 images from clinical practice. It includes a schema of 46 categories of lesion and bilingual reports. 4) **MM-Retinal** (**CFP+FFA+OCT**) contains over 4.3K high-quality image-text pairs from professional fundus diagram books, covering 96 categories of abnormalities and diseases.

Foundation Model Baselines: 1) **MoE** [28] (**MTL**) uses a multi-task approach with mixture-of-experts for fundus, macula, and optic disc images. 2) **RETFound** [35] (**MIM**) is a masked autoencoder with Transformers [7], trained for retinal image reconstruction. 3) **FLAIR** [25] (**CLIP**) utilizes a CLIP model and brief textual prompts that mapped from category labels for pre-training, but the texts are so limited and brief that fail to fully describe the images.

Evaluation Data and Metrics: For CFP modality, we evaluate on five datasets across finetuning, few-shot, and zero-shot settings, using classification accuracy (ACA) [32] for REFUGE [20], FIVES [13], ODIR200×3 [1] and TAOP [2], receiving-operating-curve(AUC) for REFUGE and AMD [9], F1-score for AMD. For FFA, we evaluate on image captioning task using the FFA-IR test split, applying BLUE 1-4, Meter, Rouge, and Cider metrics for consistency with FFA-IR.

Implementation Summary: In CFP and FFA, image and text encoders are initialized from ImageNet-1K [23] and BioClinicalBERT, respectively. Attention mechanisms are trained from scratch, with 512 feature dimensions. Images are

Table 2: Comparison of transferability on unseen downstream datasets and ablation study. FT refers to finetune.

(a) Comparison of transferability on unseen downstream datasets(CFP)										
Type	Method	Data	Data Size	FT(20%-20%)			FT(train-val)		FT(80%-20%)	
				REFUGE	FIVES	ODIR200×3	TAOP	AMD	AUC	F1
MTL	MoE	flair	278,348	0.543	0.364	0.609	0.239	0.539	0.405	0.450
MIM	RETFound	RETFound	904,170	0.809	0.765	0.907	0.697	0.945	0.805	0.821
CLIP	FLAIR	flair	278,348	0.831	0.835	0.875	0.468	0.963	0.960	0.822
		flair+syn	1,278,366	0.847	0.842	0.890	0.549	0.952	0.953	0.839
	KeepFIT	flair	278,348	0.837	0.838	0.903	0.556	0.951	0.957	0.840
		flair+syn	1,278,366	0.832	0.842	0.890	0.579	0.976	0.969	0.848
		50%flair+MM	141,343	0.856	0.834	0.913	0.700	0.962	0.962	0.871
		flair+MM	280,517	0.861	0.851	0.915	0.684	0.971	0.966	0.875
(b) Comparison of transferability on unseen downstream datasets(FFA)										
Model	Data	Data Size	B1	B2	B3	B4	Meter	Rouge	Cider	Avg
CNN + T	FFA-IR	1,048,584	0.321	0.211	0.154	0.122	0.198	0.268	0.283	0.222
CNN + T	FFA-IR+MM	1,050,531	0.363	0.244	0.171	0.127	0.149	0.302	0.314	0.239
(c) Ablation Study(MHCA denotes multi-head cross attention)										
Model	Revision	Mixed	Fusion	FT(20%-20%)			FT(train-val)		FT(80%-20%)	
	MHCA	/	MHCA	REFUGE	FIVES	ODIR200×3	TAOP	AMD	AUC	F1
KeepFIT	✓	✓		0.803	0.841	0.903	0.640	0.964	0.884	0.839
		✓		0.832	0.851	0.907	0.675	0.965	0.905	0.856
	✓	✓	✓	0.832	0.852	0.900	0.673	0.963	0.863	0.847
	(Ours) ✓	✓	✓	0.822	0.851	0.905	0.694	0.964	0.966	0.867
				0.861	0.851	0.915	0.684	0.971	0.966	0.875

adjusted to 512×512 size. All the texts we used are in English version. Text tokens length is set at 256. Evaluation uses five-fold cross-validation averaging. We employ AdamW ($lr=1e-4$, $decay=1e-5$) optimizer and cosine scheduler with initial warm-up for the first epoch. Training is conducted on 4 RTX 3090 GPUs with batches of 24.

4.2 Comparison of Generalization Ability on Zero-Shot and Few-Shot Tasks

To compare the foundational performance of different models, we conducted experiments in zero-shot and few-shot scenarios with unseen categories. We adopted 1, 5, 10 shots and adapter tuning scheme in few-shot, including ClipAdapter [10], TipAdapter [31] that adds a few task-specific parameters.

In Tab 1, KeepFIT trained by MM-Retinal and 50% flair achieves competitive performance across the board. 50% flair is obtained by randomly sampling

50% of each dataset in the flair. For example, it outperforms FLAIR trained by flair by 0.268 and KeepFIT trained by flair by 0.061 in average, and even shows superior performance to KeepFIT trained by MM-Retinal and flair with an improvement of 0.012. The results demonstrate several key insights: 1) The best performance comes from using 50% flair with MM-Retinal, which is even higher than using 100% flair with MM-Retinal. This indicates that large datasets may introduce noise and diminish transfer effectiveness since they lack expert knowledge. 2) KeepFIT performs better when trained by MM-Retinal and flair than by synfundus-1M and fliar, underscoring MM-Retinal’s superior expert knowledge over large datasets like synfundus-1M for model generalization ability and transferability.

4.3 Comparison of Transferability on Unseen Downstream Datasets

To assess KeepFIT’s transferability, we fine-tuned it on six unseen datasets. Specifically for CFP, we added and fine-tuned a fully connected layer to the image encoder and keep the other parts frozen. We conducted five fine-tuning settings, including 20%, 40%, 60%, and 80% of the data for training, the rest 20% for testing or following the official dataset partition. For FFA, due to dataset scarcity, we utilized image captioning for assessment following FFA-IR [17], using ResNet to extract images features and Transformer-based decoder for caption generation.

As shown in Tab 2(a), KeepFIT trained on MM-Retinal and flair achieves SOTA in almost all the unseen downstream datasets. Especially, it outperforms KeepFIT trained on flair by 0.035 and FLAIR trained on flair by 0.053. Similarly in Tab 2(b), the performance of FFA modality is improved when trained on both MM-Retinal and FFA-IR by 0.017. More results are provided in the supplementary material.

4.4 Ablation Study

We ablated KeepFIT on all flair and MM-Retinal from three aspects. 1) Image similarity-guided text revision module: assess the necessity of revising the text features of public datasets using the knowledge extracted from MM-Retinal by the guidance of image similarity. 2) Mixed training strategy: test the necessity of including data from two sources in one batch. 3) Text fusion module: we substituted text revision with a text fusion module. It integrates knowledge extracted from MM-Retinal by multi-head cross attention into public dataset text features via residual connections to test the necessity of augmenting text features.

Table 2(c) shows that image similarity-guided text revision module and mixed training strategy are essential for the performance improvement. This is because revision module injects the appropriate expert knowledge into the texts and mixed training strategy significantly boosts the performance by reducing conflicts. However, the text fusion module makes minimal contribution, suggesting that text revision is more effective at injecting knowledge than text fusion.

5 Conclusion

In this work, we built a multi-modal MM-Retinal dataset, with high-quality fundus image-text expertise. We also proposed KeepFIT, a vision-language pre-training framework enhancing expert knowledge infusion. Experimental results highlight its transferability to unseen datasets and generalization ability on few-shot and zero-shot scenarios. We expect this work will open up unexplored topics in fundus research, such as building multimodal knowledge graphs of fundus images, and high-quality text-to-image generation.

6 Acknowledgement

This work was partially supported by the National Natural Science Foundation of China (Grants No 62106043, 62172228), and the Natural Science Foundation of Jiangsu Province (Grants No BK20210225). We're also deeply grateful to Yu-Ang Yao, Minqi Gao, Junkai Chen, Jiaqi Li, Zimeng Zhu, and Jiaqi Xu, who have been instrumental in the construction of the MM-Retinal dataset.

References

1. International competition on ocular disease intelligent recognition (2019), <https://odir2019.grand-challenge.org>
2. Tencent miying artificial intelligence competition for medical imaging (2021), <https://contest.taop.qq.com/>
3. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323 (2019)
4. Cuadros, J., Bresnick, G.: Eyepacs: an adaptable telemedicine system for diabetic retinopathy screening. Journal of diabetes science and technology **3**(3), 509–516 (2009)
5. Decencière, E., Zhang, X., Cazuguel, G., Lay, B., Cochener, B., Trone, C., Gain, P., Ordonez, R., Massin, P., Erginay, A., et al.: Feedback on a publicly distributed image database: the messidor database. Image Analysis & Stereology **33**(3), 231–234 (2014)
6. Diao, S., Su, J., Yang, C., Zhu, W., Xiang, D., Chen, X., Peng, Q., Shi, F.: Classification and segmentation of oct images for age-related macular degeneration based on dual guidance networks. Biomedical Signal Processing and Control **84**, 104810 (2023)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)
8. Fu, H., Li, F., Orlando, J.I., Bogunović, H., Sun, X., Liao, J., Xu, Y., Zhang, S., Zhang, X.: Palm: Pathologic myopia challenge (2019). <https://doi.org/10.21227/55pk-8z03>, <https://dx.doi.org/10.21227/55pk-8z03>

9. Fu, H., Li, F., Orlando, J.I., Bogunović, H., Sun, X., Liao, J., Xu, Y., Zhang, S., Zhang, X.: Adam: Automatic detection challenge on age-related macular degeneration (2020). <https://doi.org/10.21227/dt4f-rt59>, <https://dx.doi.org/10.21227/dt4f-rt59>
10. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision **132**(2), 581–595 (2024)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
13. Jin, K., Huang, X., Zhou, J., Li, Y., Yan, Y., Sun, Y., Zhang, Q., Wang, Y., Ye, J.: Fives: A fundus image dataset for artificial intelligence based vessel segmentation. Scientific Data **9**(1), 475 (2022)
14. Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G.S., Peng, L., Webster, D.R.: Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology **125**(8), 1264–1272 (2018)
15. Lei, J., Dai, L., Jiang, H., Wu, C., Zhang, X., Zhang, Y., Yao, J., Xie, W., Zhang, Y., Li, Y., et al.: Unibrain: Universal brain mri diagnosis with hierarchical knowledge-enhanced pre-training. arXiv preprint arXiv:2309.06828 (2023)
16. 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)
17. Li, M., Cai, W., Liu, R., Weng, Y., Zhao, X., Wang, C., Chen, X., Liu, Z., Pan, C., Li, M., et al.: Ffa-ir: Towards an explainable and reliable medical report generation benchmark. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021)
18. Li, X., Zhou, Y., Wang, J., Lin, H., Zhao, J., Ding, D., Yu, W., Chen, Y.: Multi-modal multi-instance learning for retinal disease recognition. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 2474–2482 (2021)
19. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21152–21164 (2023)
20. Orlando, J.I., Fu, H., Breda, J.B., Van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.A., Kim, J., Lee, J., et al.: Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical image analysis **59**, 101570 (2020)
21. Pellegrini, C., Keicher, M., Özsoy, E., Jiraskova, P., Braren, R., Navab, N.: Xplainer: From x-ray observations to explainable zero-shot diagnosis. arXiv preprint arXiv:2303.13391 (2023)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
23. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**, 211–252 (2015)

24. Shang, F., Fu, J., Yang, Y., Huang, H., Liu, J., Ma, L.: Synfundus: A synthetic fundus images dataset with millions of samples and multi-disease annotations. arXiv preprint arXiv:2312.00377 (2023)
25. 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)
26. Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering* **6**(12), 1399–1406 (2022)
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
28. Wang, X., Ju, L., Zhao, X., Ge, Z.: Retinal abnormalities recognition using regional multitask learning. In: *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I* 22. pp. 30–38. Springer (2019)
29. Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., Duan, N.: Visual chatgpt: Talking, drawing and editing with visual foundation models. arXiv preprint arXiv:2303.04671 (2023)
30. Wu, J., Fu, R., Fang, H., Zhang, Y., Yang, Y., Xiong, H., Liu, H., Xu, Y.: Medsegdiff: Medical image segmentation with diffusion probabilistic model. In: *Medical Imaging with Deep Learning*. pp. 1623–1639. PMLR (2024)
31. Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free adaption of clip for few-shot classification. In: *European Conference on Computer Vision*. pp. 493–510. Springer (2022)
32. Zhao, Z., Zhang, K., Hao, X., Tian, J., Chua, M.C.H., Chen, L., Xu, X.: Bira-net: Bilinear attention net for diabetic retinopathy grading. In: *2019 IEEE International Conference on Image Processing (ICIP)*. pp. 1385–1389. IEEE (2019)
33. Zhou, Y., He, X., Huang, L., Liu, L., Zhu, F., Cui, S., Shao, L.: Collaborative learning of semi-supervised segmentation and classification for medical images. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 2079–2088 (2019)
34. Zhou, Y., Yang, G., Zhou, Y., Ding, D., Zhao, J.: Representation, alignment, fusion: A generic transformer-based framework for multi-modal glaucoma recognition. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 704–713. Springer (2023)
35. 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)
36. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)

Supp: MM-Retinal: Knowledge-Enhanced Foundational Pretraining with Fundus Image-Text Expertise

1 Details of MM-Retinal Dataset Construction

To advance multi-modal fundus foundation model research and foster expert knowledge integration in learning fundus image analysis models, we build MM-Retinal, a high-quality image-text paired fundus dataset that comprises CFP, FFA, and OCT modalities. We design a semi-automatic collection procedure to improve construction efficiency, which consists of four steps: 1) image-text pair collection; 2) image-text alignment; 3) modality classification; 4) text cleaning and bilingual translation.

Step 1: Image-Text Pair Collection. Textual reports for fundus diagnosis are typically not accompanied with images in clinical process. Therefore, unlike X-ray and CT images, it is challenging to directly obtain image-text pairs from fundus clinical reports. To address this, we collect image-text pairs from four diagram books illustrating ocular fundus diseases with high-quality expert captions. First, image-text pairs are captured from the books. In cases that one figure corresponds to one caption, we simply capture them in one screenshot. For the cases that multiple sub-figures correspond to a piece of caption, both the text and its corresponding sub-figures are captured together, as shown in Fig. 1(a). Note that we keep the resolution of each image in the pair no less than 800×800 . Afterwards, these screenshots are filled into the program we designed to parse images by Adobe API and extract texts by OCR technique. Additionally, the color of a certain book suffer a dark tone. We implemented a dehazing operation based on Gamma transformation to correct the color of this book to match the color distribution of other books.

Step 2: Image-Text Alignment. As mentioned earlier, not all the initially extracted images and texts by the program are very well aligned, as there are some sub-figures correspond to one caption within a screenshot. Thus, for those failure cases, we use regular expression matching to split them. Specifically, if the text matches "Figure No.", it indicates the beginning of the text of a new image-text pair. If it matches "Letter.", it indicates the beginning of the text of a new subfigure-text pair. As for the separation of sub-figures, we apply Adobe API to automatically implement subgraph segmentation.

Step 3: Modality Classification. Since the fundus images in the books include multiple modalities, we classified them into CFP, FFA, and OCT, separately, which are the mainstream modalities in fundus domain, and exclude others due to their limited samples. Specifically, we first employ K-means to categorize images into three categories based on their color histograms. It effectively

separates the CFP modality which characterized by the distinct color pattern. The other two categories are the mixture of FFA, OCT and images from other modalities. Subsequently, we select a reference image from both OCT and FFA respectively. The classification of remaining images is based on the distance between their color histograms and the references' color histograms. This process results in a precise separation of FFA and OCT modality images.

Step 4: Text Cleaning and Bilingual Translation. The text extraction may have some OCR recognition errors or incomplete issues, so we correct the texts manually. Moreover, to enhance the relevance within an image-text pair, we remove irrelevant information from the text, such as the index of the corresponding image. We also make modifications to sentence inaccuracies. For example, the original text describes multi subgraphs (e.g. ‘both eyes show’), but after separating them, the text description should be appropriately adjusted to reflect details about each individual subgraph (e.g. ‘left eye shows’ and ‘right eye shows’). In addition, since the diagram books we selected include both Chinese and English, we provide bilingual reports in English and Chinese version to standardize the language and make MM-Retinal more influential. DeepL Translator is applied to translate Chinese to English, and Tencent Translator is used to translate English to Chinese.

2 Few-Shot on Unseen Downstream Datasets

In this section, we present the supplementary results of Few-Shot experiments. The metric used for REFUGE, ODIR200 \times 3 and FIVES is ACA, and for AMD is AUC. From the Fig. 1, it is evident that KeepFit performs better with smaller training splits, like 20%-60%. As the training data increases, KeepFit’s performance slightly declines, yet it consistently outperforms most baselines. This demonstrates its superior generalization capabilities and robustness.

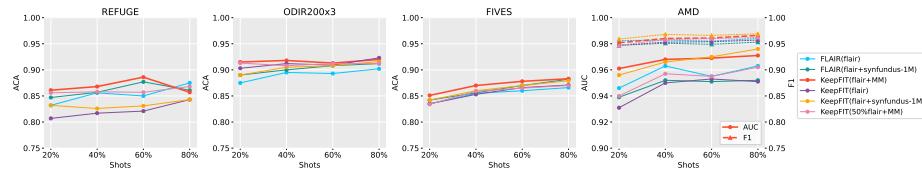


Fig. 1: Few-shot experiments on unseen downstream datasets.

3 Disease and Abnormal Changes Categories in MM-Retinal

This section provides the major categories of fundus diseases and abnormal changes in MM-Retinal with several example images in CFP, FFA and OCT

modalities. These categories are summarized from the contents of four diagram books we used.

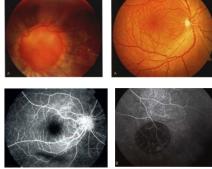
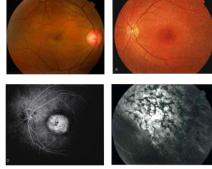
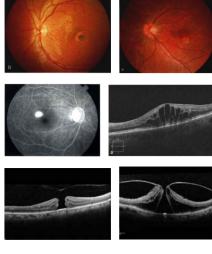
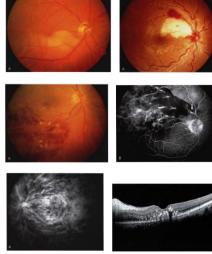
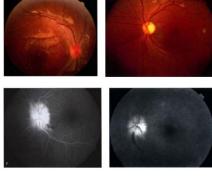
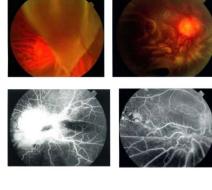
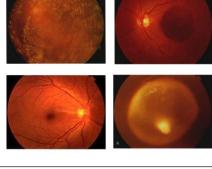
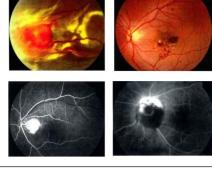
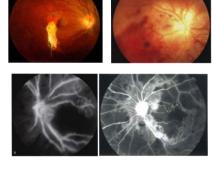
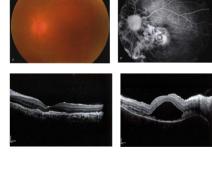
Disease of congenital anomalies 1) Congenital anomalies of vitreous 2) Abnormal development of optic nerve 3) Abnormal development of retina and choroid 4) Congenital macular anomalies 5) Congenital retinal vascular anomalies 6) Congenital choroidal defect		Chorio-retinal degeneration and dystrophy 1) Retinal degeneration and dystrophy 2) Cone dystrophy 3) Pattern dystrophy 4) Congenital stationary night blindness 5) Choroidal dystrophies	
Macular diseases 1) Macular degeneration and dystrophy 2) Degenerative myopia 3) Central serous chorioretinopathy 4) Multiple posterior pole retinal pigment epitheliopathies 5) Occult macular dystrophy 6) High myopia macular degeneration 7) Central exudative chorioretinopathy 8) Age-related macular degeneration 9) Idiopathic polypoidal choroidal vasculopathy 10) Vitreomacular traction syndrome 11) Choroidal Folds		Retinal vascular diseases 1) Retinal vascular occlusive diseases 2) Hypertensive retinopathy 3) Diabetic retinopathy 4) Retinal vasculitis 5) Complications of vascular diseases 6) Cilioretinal artery obstruction 7) Sickle cell retinopathy 8) Lipemia Retinalis 9) Radiation Retinopathy 10) Ocular ischemic syndrome 11) Other retinal vascular diseases	
Optic nerve diseases 1) Optic neuritis 2) Optic disc vasculitis 3) Ischemic optic neuropathy 4) Multiple sclerosis 5) Neuromyelitis optica 6) Leber's hereditary optic neuropathy 7) Optic atrophy 8) Optic disk edema 9) Optic disk vitreous warts 10) Pseudopapchia		Retinal and choroidal detachment 1) Rhegmatogenous retinal detachment 2) Special type retinal detachment 3) Traction retinal detachment 4) Exudative retinal detachment 5) Traumatic retinal detachment 6) Bullous retinal detachment 7) Proliferative vitreoretinopathy 8) Rhegmatogenous retinal detachment	
Disease of vitreous 1) Asteroid hyalosis 2) Synchysis scintillans 3) Persistent hyperplasia of primary vitreous 4) Congenital cyst of vitreous 5) Vitreous parasite		Intraocular tumors 1) Tumors of Optic disc 2) Tumors of retina 3) Tumors of choroid 4) Ciliary body melanoma 5) Ocular Lymphoma	
Traumatic and toxic retinopathies 1) Solar maculopathy 2) Traumatic macular hole 3) Chorioretinitis scleropatia 4) Talc retinopathy 5) Chloroquine or hydroxychloroquine retinopathy 6) Thioridazine retinopathy		Retinal and choroidal inflammatory diseases 1) Infectious chorioretinitis 2) Noninfectious chorioretinitis 3) Other choroidal diseases	

Fig. 2: Major fundus diseases in MM-Retinal

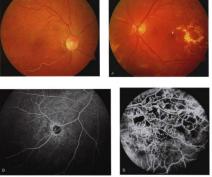
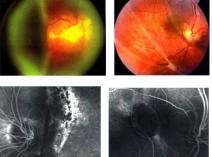
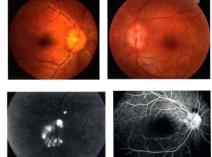
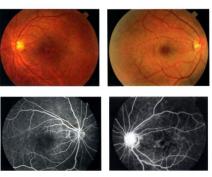
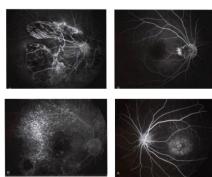
Retinal vascular		Vitreous	
1) Commotio Retinæ 2) Retinal vascular changes 3) Retinal vascular hemorrhages 4) Retinal exudates and hemorrhages 5) Retinal drusens 6) Retinal breaks 7) Retinal detachment 8) Gyrate atrophy		1) Vitreous amyloidosis 2) Vitreous haemorrhages 3) Posterior vitreous detachment 4) Avulsed Vitreous Base	
Choroidal		Optic disk	
1) Choroidal detachment 2) Choroidal rupture		1) Papilledema 2) Optic disc drusen 3) Fundus changes in hypermyopia 4) Fundus changes in high blood pressure and atherosclerosis	
Macular		Abnormal fluorescence in angloaram	
1) Subretinal neovascularization 2) Macular hole 3) Macular cystoid edema		1) Hyperfluorescence	
Crystalline lens			
1) Dislocated Lens			

Fig. 3: Major abnormal fundus changes in MM-Retinal