# A Foundation LAnguage-Image model of the Retina (FLAIR): Encoding expert knowledge in text supervision

Julio Silva-Rodriguez[a], Hadi Chakor[b], Riadh Kobbi[b], Jose Dolz[a,c], Ismail Ben Ayed[a,c]

[a]*ETS Montreal, Quebec, Canada*
[b]*DIAGNOS Inc., Quebec, Canada*
[c]*Centre de Recherche du Centre Hospitalier de l'Université de Montréal (CR-CHUM), Canada*

**Abstract**

Foundation vision-language models are currently transforming computer vision, and are on the rise in medical imaging fueled by their very promising generalization capabilities. However, the initial attempts to transfer this new paradigm to medical imaging have shown less impressive performances than those observed in other domains, due to the significant domain shift and the complex, expert domain knowledge inherent to medical-imaging tasks. Motivated by the need for domain-expert foundation models, we present FLAIR, a pre-trained vision-language model for universal retinal fundus image understanding. To this end, we compiled 37 open-access, mostly categorical fundus imaging datasets from various sources, with up to 97 different target conditions and 284,660 images. We integrate the expert's domain knowledge in the form of descriptive textual prompts, during both pre-training and zero-shot inference, enhancing the less-informative categorical supervision of the data. Such a textual expert's knowledge, which we compiled from the relevant clinical literature and community standards, describes the fine-grained features of the pathologies as well as the hierarchies and dependencies between them. We report comprehensive evaluations, which illustrate the benefit of integrating expert knowledge and the strong generalization capabilities of FLAIR under difficult scenarios with domain shifts or unseen categories. When adapted with a lightweight linear probe, FLAIR outperforms fully-trained, dataset-focused models, more so in the few-shot regimes. Interestingly, FLAIR outperforms by a large margin more generalist, larger-scale image-language models, which emphasizes the potential of embedding experts' domain knowledge and the limitations of generalist models in medical imaging. The pre-trained model is available at: `https://github.com/jusiro/FLAIR`.

*Keywords:* Foundation models, Fundus image analysis, Vision-language pre-training, Expert knowledge.

## 1. Introduction

At least 1 billion people have a vision impairment that could have been prevented or is yet to be addressed (WHO, 2019). In this context, fundus color images combined with computer vision systems present a promising, cost-effective solution for large-scale screening and early detection of ophthalmologic diseases (Balyen and Peto, 2019; Bellemo et al., 2019).

Driven by public datasets, deep learning has reached remarkable performances in a breadth of fundus image analysis problems, such as diabetic retinopathy grading (Liu et al., 2022), glaucoma detection (Orlando et al., 2019), lesion segmentation (Porwal et al., 2020) or multi-disease detection (Cen et al., 2021). Nevertheless, several limitations impede the widespread adoption of these methods. In particular, current deep learning solutions for fundus image analysis may not generalize well whenever there are shifts in the imaging data or in the task at hand (e.g., new or rare classes) (Li et al., 2021; Sengupta et al., 2020). In retinal imaging, and in the much broader field of medical imaging, the current dominant deep learning paradigm is to supervise models on very specific tasks, e.g., diabetic retinopathy classification into a few grades (Liu et al., 2022). Learning representations that might be too specialized for the task and training images at hand, such task-focused models may have difficulty in (i) dealing with the high variability in real clinical scenarios (Finlayson et al., 2021), due to the high variations in image acquisition and patient demographics; and (ii) capturing rare conditions that are

*Corresponding author: julio-jose.silva-rodriguez@etsmtl.ca

not well represented in the training data.

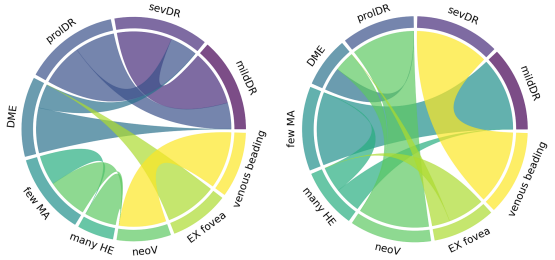
There is currently a paradigm shift in artificial intelligence algorithms, driven by the growing prevalence of models trained on large and diverse datasets, which could be adapted to a broad span of downstream tasks. These models, commonly referred to as *foundation* models, have gained increasing popularity and showed significant success in computer vision and natural language processing tasks (Brown et al., 2020; Radford et al., 2021). In particular, vision-language models such as CLIP (Radford et al., 2021) or ALIGN (Jia et al., 2021) have shown impressive generalization capabilities when fine-tuned on various downstream computer-vision tasks, emerging as powerful alternatives to narrowly-supervised, task-focused models. Learning from large-scale amounts of image-text pairs, such models leverage the rich semantic knowledge in the language-based supervision, thereby yielding visual features that are more descriptive than their task-specific counterparts.

In computer vision tasks, such as image classification, this new *pretrain-and-finetune* paradigm enhanced robustness to image-data shifts and showed promising zero-shot and few-shot transferability. Nonetheless, initial attempts to directly apply these foundation models to the medical domain yielded less convincing performances (Wang et al., 2022b). Indeed, generalist models like CLIP may not capture the fine-grained image features and class dependencies/hierarchies, which might be complex, highly specialized concepts inherent to the expert's domain knowledge; see Figure 1 for an illustration in the case of retinal fundus images. This has recently motivated the development of foundation models specialized for medical imaging applications (Wójcik, 2022; Moor et al., 2023).

Vision-language models are currently emerging in medical image analysis. Several recent studies investigated foundation models specialized in radiology (Zhang et al., 2022b; Huang et al., 2021b; Wang et al., 2022b), focusing mostly on chest-radiography data. These were motivated by the prevalence of diagnostic text reports in radiology, and the availability of large domain resources to mine such textual information (Bodenreider, 2004; Jain et al., 2021). However, this may not be the case in other medical-imaging modalities. In retinal imaging, for instance, text information is scarce and most datasets are categorically labeled (see Table 1), i.e., the label of each training image is a single category (or class), e.g., "*mild diabetic retinopathy*" (mildDR).

We argue that, even for categorically-labeled images, vision-language pre-training is an appealing solution to integrate domain-specific, fine-grained knowledge, such as the dependencies between the categories, into visual representations. The analysis of medical images by clinical experts is a process of searching for differential features of candidate conditions. In this process, there are, for instance, hierarchical dependencies between the presence of local lesions and the differential diagnosis at the global level. Such expert's domain knowledge is usually overlooked in conventional training but could be integrated in the form of text descriptions, to build powerful image-language models. To illustrate this, we provide in Figure 2 a few retinal-imaging examples with categorical labels along with the corresponding text descriptions encoding domain knowledge. For instance, the text description "*only a few microaneurysms are present*" informs on local conditions known to point to the category mildDR (Wilkinson et al., 2003). In Table S3, we provide a comprehensive list of the correspondences between the categorical labels and textual domain-knowledge descriptions, which we compiled from the relevant clinical literature (Garner and Ashton, 1979) and from community standards (Wilkinson et al., 2003), to build our foundation model of the retina.

In this work, we introduce FLAIR, a Foundation LAnguage-Image model of the Retina, for fundus image analysis. FLAIR is trained and validated on a large assembly of 37 datasets, with 284,660 images and 96 different target categories, which we compiled from different publicly available sources. We integrate the expert's domain knowledge in the form of text supervision dur-




(a) CLIP (Radford et al., 2021) (b) FLAIR (*Ours*)

Figure 1: **CLIP limitations on medical domains.** The figure depicts the cosine similarities of the text embeddings for common retinal diseases and lesions observed on fundus images. While CLIP mostly focuses on general medical relations (e.g., "*diabetic*", or "*neovascularization*"- "*venous*"), the proposed domain-specific model (i.e., FLAIR) is able to capture the hierarchical dependencies between concepts (e.g., the fundus images of "*mildDR*" contain "*only a few microaneurysms*", and "*neovascularization*" is the differential sign for "*prolDR*" diagnosis).

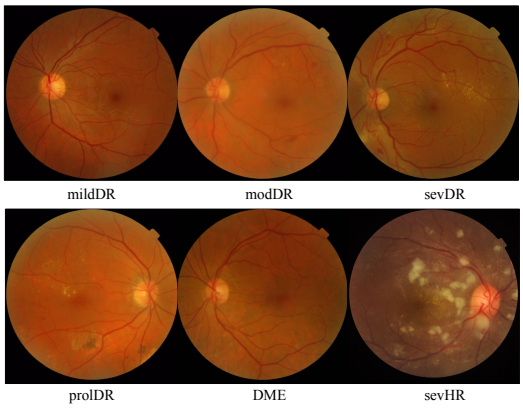



Figure 2: **Expert knowledge descriptors.** The analysis of fundus images by ophthalmologists is driven by hierarchical features. According to the American Academy of Ophthalmology (Wilkinson et al., 2003), mildDR is characterized by "*only few microaneurysms present*", modDR includes "*retinal haemorrhages in few quadrants*", "*many haemorrhages*" or "*cotton wool spots*", and sevDR and prolDR are distinguished by "*venous beading*"/"*intraretinal microvascular abnormalities*" and "*neovascularization*", respectively. DME is also usually featured by "*hard exudates involving the center of the macula*". Furthermore, according to (Garner and Ashton, 1979), hypertensive retinopathy is generally described as "*flame-shaped hemorrhages in the superficial layers of the retina and cotton-wool patches*". Going deeper into the hierarchies between concepts, exudates are "*small white or yellowish deposits*", and microaneurysms are "*small red dots*".

ing both pre-training and zero-shot prediction, thereby enhancing the categorical information of the data. Such a textual expert's knowledge describes the fine-grained features of the pathologies as well as the hierarchies and dependencies between them. We report comprehensive evaluations, comparisons and ablation studies, which show the substantial effect of embedding expert knowledge and the strong generalization and transferability capabilities of FLAIR under challenging scenarios with domain shifts or novel (unseen) categories. When adapted with a lightweight linear probe classifier, FLAIR outperforms models that are fully trained on the target dataset, more so under low-data (few-shot) settings. Furthermore, FLAIR outperforms by a large margin more generalist, larger-scale image-language models such as CLIP or BiomedCLIP. Our results point to the potential of embedding expert domain knowledge and to the limitations of generalist models.

## 2. Related Work

### *2.1. Transfer learning in medical image analysis*

Training robust deep-learning models from scratch requires large datasets and huge computational resources (Erhan et al., 2009). These conditions are rarely met in medical imaging. The high variability in image acquisition, the low prevalence of certain conditions, and the limited resources of institutions make it difficult for standard supervised-learning models to capture the substantial variability in real clinical contexts. This is due to the fact that supervised models are typically trained on and specialized for relatively small data sets and tasks, due to the prohibitive costs and resources of labeling the data. To mitigate this to some extent, transfer learning from natural images, whereby a deep model is pre-trained on a large labeled dataset such as ImageNet and then used as initialization for adaptation to a target task, has become the *de-facto* solution (Raghu et al., 2019; Kanavati and Tsuneki, 2021; Matsoukas et al., 2022). In particular, fine-tuning the whole network, or just the last layers, demonstrated promising performances in a breadth of medical-imaging tasks, across various domains such as radiology, cardiology, and ophthalmology (Tajbakhsh et al., 2017; Abramoff et al., 2016; Fauw et al., 2018). Nonetheless, several in-depth empirical studies have exposed the limited performance gains of such transfer-learning solutions in certain scenarios in medical-image classification (Raghu et al., 2019; Neyshabur et al., 2020; Matsoukas et al., 2022). Larger-scale pre-training, which leverages unlabeled data via self-supervision (Chen et al., 2022a; Huang et al., 2023), is a promising alternative. However, supervised, domain-specific pre-training remains the prevalent solution for optimal transfer learning (Zhang et al., 2022b; Liu et al., 2023). Under this standard supervised-learning paradigm, top-competing solutions on the DeepDRiD challenge for diabetic retinopathy grading on fundus images (Liu et al., 2022) performed an exhaustive, task-specific pre-training using public datasets.

### *2.2. From supervised, task-specific models to large-scale vision-language pre-training*

As discussed above, supervised, task-specific models are currently prevalent in medical imaging. Nevertheless, the generalization of such task-focused models across the various conditions encountered in clinical scenarios remains a major challenge for wider adoption (Finlayson et al., 2021). Task-specific models, e.g., diabetic retinopathy classification into a few grades, might yield features that are too specialized for the task and training data at hand. As we will show in the empirical validation of this work, such models generalize poorly whenever there are shifts in the task (e.g., new classes) or in the imaging data. Such shifts occur frequently in

medical imaging due to the high variability in image acquisition and patient demographics, and/or to the low prevalence of certain conditions.

Very recently, Vision-Language Pre-training (VLP) has made substantial progress in computer vision and machine learning, emerging as a powerful solution to improve the generalization of deep models. Large-scale VLP models leverage paired image-language data through image-text contrastive learning (Radford et al., 2021; Jia et al., 2021; Yang et al., 2022), yielding robust and generic feature extractors. Such pre-training models have shown impressive transfer-learning capabilities when fine-tuned on downstream tasks (Radford et al., 2021). On natural images, this widely emerging pre-train-and-finetune paradigm yielded excellent robustness to data shifts and strong generalization to new tasks, with no (i.e., zero-shot (Shu et al., 2022; Zhao et al., 2022)) or only a few (i.e., few-shot (Hu et al., 2022)) labeled samples in the new task. For instance, CLIP (Radford et al., 2021) provides prompt-based (zero-shot) classifications by capturing the similarity between the image and a textual description of the target class, via the jointly trained vision and language encoders. In addition, in few-shot regimes (i.e., when few labeled samples in the target task are available), the visual representations show strong transferability by updating a linear-classifier layer on top of the frozen network. Such a fast fine-tuning procedure is commonly referred to as Linear Probing (LP). These observations have motivated an increasing interest in efficient forms of adapting CLIP to downstream tasks and domains, using lightweight multi-modal modules, known as adapters (Gao et al., 2021; Zhang et al., 2022a). Although these efficient transferability properties are of huge interest in medical imaging analysis, applying CLIP directly to medical imaging data yields sub-optimal results (Wang et al., 2022b). Along with the domain shifts occurring in the imaging modality, this might be due, in part, to the complex, highly specialized terminology encountered in medical imaging.

### *2.3. Vision-language models in medical imaging*

The rise of VLP models is at its beginning in medical imaging. Several recent works investigated contrastive image-language models tailored to medical data Zhang et al. (2022b); Huang et al. (2021b); Wang et al. (2022b); Lu et al. (2023), but mostly in the application area of chest radiographs. VLP models are particularly appealing in the field of radiology (Zhang et al., 2022b), as the diagnostic text reports associated with the images are common in everyday radiology practices. Thus, public, large-scale, and multi-modal datasets with paired images and language descriptions started to appear recently, such as MIMIC-CXR (Johnson et al., 2019), PadChest (Bustos et al., 2019) or ROCO (Pelka et al., 2018). In addition, there exist large domain resources, such as UMLS (Bodenreider, 2004), BioClinicalBERT or RadGraph (Jain et al., 2021), which favor the processing and knowledge extraction of structured clinical information from free-text radiology reports.

Fueled by the existence of this domain knowledge, several recent works have developed strategies to overcome the limitations of CLIP in the medical field. For example, methods using CLIP for inference have integrated modality prompts (Liu et al., 2023), or attribute descriptions (Menon and Vondrick, 2023) for prompt-based inference, using pre-trained question-answering models to describe the shape and color of the target conditions (Qin et al., 2023). Other works have focused on the pre-training stage, generating domain-specialized VLP models such as ConVirt (Zhang et al., 2022b), PubMedCLIP (Eslami et al., 2021), GLorIA (Huang et al., 2021b), MedCLIP (Wang et al., 2022b) or MedKLIP (Wu et al., 2023), among others (Windsor et al., 2023; Wang et al., 2022a; Müller et al., 2022; Chen et al., 2022b). One of the main challenges of such pre-training lies in the low prevalence of text-based supervision on publicly available datasets. To alleviate this issue, MedCLIP incorporated categorically-labeled samples through label-space alignment (Wang et al., 2022b). Other methods have taken profit from well-established domain tools in radiology such as UMLS and RadGraph to augment the available text reports (Chen et al., 2022b; Wu et al., 2023).

Despite these recent advances in the development of vision-language pre-training strategies in medical imaging, the use of categorically-labeled datasets has been overlooked. In this work, we argue and show that such a categorical supervision could still be exploited to train powerful vision-language representations, by encoding expert's domain knowledge into text supervision.

### *2.4. Expert knowledge-driven models of fundus images*

The general idea of integrating domain knowledge into deep learning for medical image analysis is not new, and has triggered interest in the recent literature (Xie et al., 2021). In particular, domain-specific, expert knowledge (EK) from clinicians could be retrieved to highlight areas of interest, relevant features, anatomical priors, or inter-disease dependencies and hierarchies. In the case of retinal imaging, the expert's knowledge has been integrated in various ways. For instance, Giancardo et al. (2012) first segmented the exudates, which served as a proxy for macular edema detection.

Similarly, several other strategies train attention modules to enhance local lesions, which act as surrogates for disease classification. Closely related to our work, we have identified several categories, which include: using pixel-level annotated lesions for AMD staging (Fang et al., 2019), weakly-supervised strategies based on the relationships between diabetic retinopathy and diabetic macular edema (Xiaomeng et al., 2020), or disentangling disease-specific saliency maps for diabetic retinopathy grading (Sun et al., 2021). In addition, expert knowledge for glaucoma detection in fundus images is usually integrated by cropping the optic-disk area as an initial step before classification (Diaz-Pinto et al., 2019; de Vente et al., 2023). Unlike this existing literature, we study the use of well-established expert knowledge on retinal image analysis via vision-language pre-training, which has been largely overlooked in the context of foundation models. Concretely, we propose a contrastive image-text pre-training, which incorporates relevant features, hierarchies, and relationships between the classes as well as information on the regions of interest characterizing the target diseases, in the form of descriptive textual prompts, paired with the corresponding images.

## 3. Methodology

Fig. 3 depicts an overview of our framework. We introduce each methodological component formally in the following sections.

***Problem setup***. Let us define an assembly dataset, $\mathcal{D}_T$, which contains $N$ samples gathered from different publicly available fundus image datasets, including heterogeneous sources and findings. For each sample, we build a multi-modal triplet including an image, a categorical label and a text description: $\mathcal{D}_T = \{(\mathbf{X}_n, y_n, \mathbf{T}_n)\}_{n=1}^{N}$. $\mathbf{X}_n \in \mathbb{R}^{\Omega_n}$ denotes a fundus 2D image, with $\Omega_n$ its spatial domain, $y_n \in \{1, ..., C\}$ is a label among the $C$ unique categories in the assembly dataset, and $\mathbf{T}_n \in \mathcal{T}$ is a text description associated with the label. Figure 2 provides a few examples of categorical labels, such as DME, and the associated text descriptions encoding domain knowledge, e.g., "*hard exudates involving the center of the macula*". Such textual domain knowledge could be derived from the relevant clinical literature (Garner and Ashton, 1979) and/or from community standards (Wilkinson et al., 2003). Table S3 provides a comprehensive list of the correspondences between the categorical labels and textual domain-knowledge descriptions, which we compiled from the relevant clinical literature, to build our foundation model of the retina. Note that a single categorical label may correspond to several text descriptions, each describing a different finding or feature in the image. The objective of our vision-language pre-training is to provide a powerful multi-modal model capable of learning a feature representation space where samples are aligned across the three modalities: images, categories, and text.

### *3.1. Aligning images, labels and domain-knowledge text*

Our multi-modal pre-training integrates vision and language encoders. Let $\theta = \{\theta_f(\cdot), \theta_p(\cdot)\}$ denotes the vision encoder, with $\theta_f(\cdot)$ a feature extractor and $\theta_p(\cdot)$ a projection head. The feature extractor $\theta_f(\cdot)$ yields a feature representation $\tilde{\mathbf{u}} \in \mathbb{R}^{D_\mathbf{u}} : \tilde{\mathbf{u}}_i = \theta_f(\mathbf{X}_i)$ of an input image $\mathbf{X}_i$, with $D_\mathbf{u}$ the dimension of the visual feature space. Analogously, let $\phi = \{\phi_f(\cdot), \phi_p(\cdot)\}$ denotes the text encoder, $\phi_f(\cdot)$ being a feature extractor and $\phi_p(\cdot)$ a projection head. The feature extractor $\phi_f(\cdot)$ provides an embedding $\tilde{\mathbf{v}} \in \mathbb{R}^{D_\mathbf{v}} : \tilde{\mathbf{v}}_j = \phi_f(\mathbf{T}_j)$ of an input text $\mathbf{T}_j$, with $D_\mathbf{v}$ denoting the dimension of the space of text features. Each of the projection heads, $\theta_p(\cdot)$ and $\phi_p(\cdot)$, maps the independent modality representations into a joint unit hyper-sphere space: $\mathbf{u} = \frac{\theta_p(\tilde{\mathbf{u}})}{\|\theta_p(\tilde{\mathbf{u}})\|}$ and $\mathbf{v} = \frac{\phi_p(\tilde{\mathbf{v}})}{\|\phi_p(\tilde{\mathbf{v}})\|}$. In this normalized space, the similarity between image $\mathbf{X}_i$ and text description $\mathbf{T}_j$ is evaluated by the cosine similarity: $\mathbf{u}_i^T \mathbf{v}_j$, where $T$ denotes the transpose operator.

The objective consists in learning feature representations that minimize the distances between paired image and text descriptions while maximizing the distances between unpaired samples. We build image-text pairs from the available categorical label information, thereby encouraging samples belonging to the same category to have close feature representations, in both the image and text domains. More formally, let $\mathcal{B}$ denote a batch containing a set of images $\{\mathbf{X}_i\}_{i \in \mathcal{X}_B}$ and a set of text descriptions $\{\mathbf{T}_j\}_{j \in \mathcal{T}_B}$, where $\mathcal{X}_B \subset \{1, \ldots, N\}$ denotes the set of indices of the images in $\mathcal{B}$, and $\mathcal{T}_B \subset \{1, \ldots, N\}$ the set of indices of the text descriptions in $\mathcal{B}$. We minimize category-aware image-to-text ($\mathcal{L}_{i2t}$) and text-to-image ($\mathcal{L}_{t2i}$) contrastive objectives, defined as follows:

$$\mathcal{L}_{i2t}(\theta, \phi, \tau|\mathcal{B}) = -\sum_{i \in \mathcal{X}_B} \frac{1}{|P_{\mathcal{T}_B}(i)|} \sum_{i' \in P_{\mathcal{T}_B}(i)} \log \frac{\exp(\tau \mathbf{u}_i^T \mathbf{v}_{i'})}{\sum_{j \in \mathcal{T}_B} \exp(\tau \mathbf{u}_i^T \mathbf{v}_j)} \tag{1}$$

$$\mathcal{L}_{t2i}(\theta, \phi, \tau|\mathcal{B}) = -\sum_{j \in \mathcal{T}_B} \frac{1}{|P_{\mathcal{X}_B}(j)|} \sum_{j' \in P_{\mathcal{X}_B}(j)} \log \frac{\exp(\tau \mathbf{u}_{j'}^T \mathbf{v}_j)}{\sum_{i \in \mathcal{X}_B} \exp(\tau \mathbf{u}_i^T \mathbf{v}_j)} \tag{2}$$

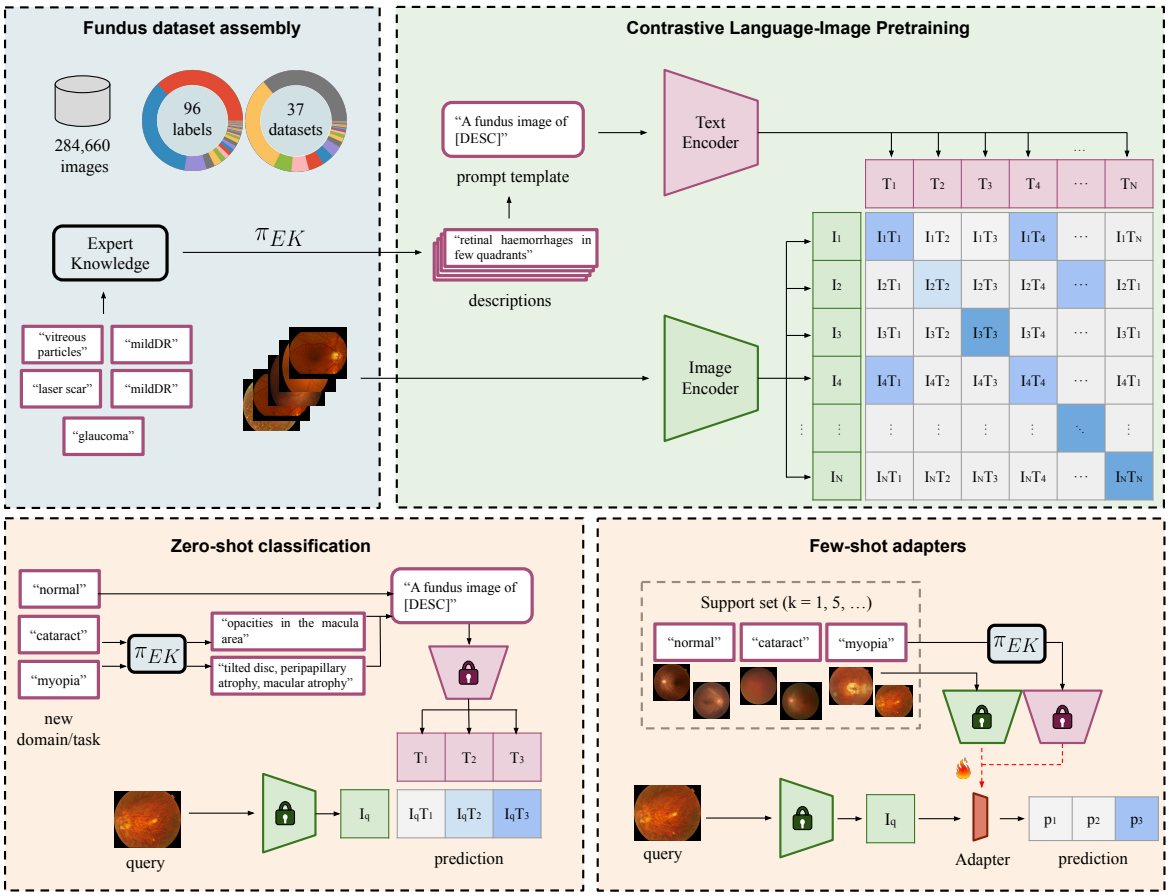



Figure 3: **Framework overview**. We have developed a knowledge-based universal model of the retina from an assembly of 37 public datasets, which contains 284,660 fundus images and 96 different categories (*see top-left*). The foundation model consists of vision and language encoders, which are trained in a contrastive fashion on paired images and textual descriptors. To mitigate the scarcity of text-based supervision in publicly available retinal fundus imaging datasets, we propose to augment the categorical image labels by using well-established domain knowledge (*see top-right*). The ensuing pre-training model enables to predict new categories in a zero-shot fashion, using well-designed descriptors based on domain knowledge and local features of the novel diseases; *see bottom-left*. In addition, the model could adapt to downstream tasks and domains by tuning a lightweight adapter on top of the image and vision encoders, by using only a few labeled samples (the support set); *see bottom-right*.

where $\tau \in \mathbb{R}_{>0}$ is a trainable scaling parameter, $|\cdot|$ denotes the cardinality of a set and $P_{\mathcal{T}_B}(i)$ and $P_{\mathcal{X}_B}(j)$ contain indices of similar-category subsets of batch $\mathcal{B}$:

$$P_{\mathcal{T}_B}(i) = \{i' | i' \in \mathcal{T}_B, y_{i'} = y_i\} \text{ and } P_{\mathcal{X}_B}(j) = \{j' | j' \in \mathcal{X}_B, y_{j'} = y_j\}$$

Thus, the vision and language encoders of the foundation fundus model are trained to optimize a bidirectional image and language alignment, using gradient descent and randomly batched samples:

$$\min_{\theta,\phi,\tau} \sum_{\mathcal{B}} \mathcal{L}_{i2t}(\theta, \phi, \tau|\mathcal{B}) + \mathcal{L}_{t2i}(\theta, \phi, \tau|\mathcal{B}) \quad (3)$$

### 3.2. *Expert knowledge as additional text supervision during pre-training.*

Large-scale vision-language pre-training for medical imaging has been mostly investigated in the context of radiology images (Zhang et al., 2022b), where diagnostic text reports are common. Thus, large multi-modal datasets with paired radiology images and text descriptions are available (Johnson et al., 2019). Nevertheless, this is not the case in other medical imaging modalities, such as retinal fundus images. In the vast majority of the datasets in our assembly data $\mathcal{D}_T = \{(\mathbf{X}_n, y_n, \mathbf{T}_n)\}_{n=1}^N$, the text representation $\mathbf{T}_n$ is not available (see Table 1). Therefore, we introduce a mapping function, $\pi(\cdot) : \mathcal{Y} \to \mathcal{T}$, which generates text descriptions from the categorical labels, thereby building our multi-modal dataset as $\mathcal{D}_T = \{(\mathbf{X}_n, y_n, \pi(y_n))\}_{n=1}^N$.

A typical solution would be to use a *naive* transfer function, $\pi_{naive}(\cdot)$, which only brings information on the imaging modality used (e.g. (Liu et al., 2023) for CT volumes). In the case of fundus images, the modality prompt template would thereby be "*A fundus photograph of [CLS]*", where "*[CLS]*" indicates the category name. Although this solution might integrate semantic relations of similarly-named categories, it fails to capture domain-specific hierarchies and thus may become *uninformative* (Menon and Vondrick, 2023). In this work, we propose to exploit well-established *domain expert knowledge* (EK) descriptions, which we denote as $\pi_{EK}(\cdot)$. This transformation maps each category to text descriptions that contain relevant findings for each disease, as well as inter-category relationships. Given a category $y^*$, the mapping produces an ensemble of $P$ text descriptions such that $\{\mathbf{T}^*\}_1^P = \pi_{EK}(y^*)$. For example, a text description of category "*no diabetic retinopathy*" would be "*no relevant haemorrhages, microaneurysms or exudates"*, while the category "*exudates*" could be described as "*small white or yellowish-white deposits with sharp margins*". It is worth mentioning that $P$, the number of textual expert knowledge descriptions, might differ from one category to another. For additional examples, we refer the reader to Table S3 in Supplemental Materials. Thus, during the stochastic gradient descent optimization of the loss function in 3, and for a given sample of a categorically-labeled dataset in a batch $\mathcal{B}$, a text description is uniformly sampled from a set containing the naive prompt and the corresponding expert knowledge prompts. By encoding this domain knowledge during training, large-scale vision-language pre-training can capture stronger inter-category relations, potentially leading to richer representations.

### *3.3. Expert knowledge to enhance zero-shot inference.*

Vision-language pre-trained models might serve as a powerful tool for zero-shot classification. This involves the generalization to *unseen datasets*, which might present novel target tasks. Formally, let us define a target image, $x^*$, and a given set of novel categories within the target dataset, $y' \in \{C+1, ..., C'\}$. In this case, the inference is driven by the observed cosine similarity between the image representation produced by the image encoder, $\mathbf{u}^*$, and a representation of each category produced by the text encoder, $\mathbf{v}_{c'}$, using a language description of each category, $\pi(\{c'\}_{C+1}^{C'})$. Specifically, the predicted category corresponds to the maximum cosine similarity:

$$\underset{c'}{argmax} \frac{\exp(\tau \mathbf{u}^{*T} \mathbf{v}_{c'})}{\sum_{c \in \{C+1,...,C'\}} \exp(\tau \mathbf{u}^{*T} \mathbf{v}_c)} \tag{4}$$

Regarding the text representation, the first works using CLIP-alike models for inference resorted to a naive prompting strategy based on category name Radford et al. (2021). Nevertheless, recent reports indicate that the category names might overlook the full value of the additional information provided by the language modality (Menon and Vondrick, 2023). For instance, (Menon and Vondrick, 2023) investigated how to enhance class representations using large language models in the context of natural images, and (Wang et al., 2022b; Wu et al., 2023) used domain-knowledge prompts in the context of radiology images. In line with these recent developments, we take advantage of the expert-knowledge prompts during the inference phase, in addition to their use during training, which we introduced in the previous sub-sections. This could potentially enhance discrimination in pathologies that are not seen during training, based on the description of their underlying characteristics. In this setting, for each novel category $c$, we compute its textual representation $\mathbf{v}_c$ in Eq. 4 as the centroid of the $P$ text embeddings of the expert-knowledge prompts corresponding to category $c$.

## 4. Experimental setting

### *4.1. Datasets*

***Assemblying the dataset*** $\mathcal{D}_T$. A total of 37 public datasets are assembled for training and evaluating the proposed universal model. A summary of the datasets is presented in Table 1. The assembled dataset combines the main tasks explored for fundus image analysis, which include: diabetic retinopathy grading (Decencière et al., 2014; Porwal et al., 2020; Castillo Benítez et al., 2021; Takahashi et al., 2017; Lin et al., 2020; Li et al., 2019b; Nakayama et al., 2023), glaucoma detection (Li et al., 2019a; Kovalyk et al., 2022; Diaz-Pinto et al., 2019; Kumar et al., 2023; de Vente et al., 2023; Orlando et al., 2019) and lesion segmentation (Pires et al., 2014; Lin et al., 2020; Li et al., 2019b; Kauppi et al., 2007; Giancardo et al., 2012). Furthermore, we included datasets that target the classification of other, less-prevalent diseases (Pachade et al., 2021; Cen et al., 2021; Nakayama et al., 2023; Hassan et al., 2021). While most of the datasets contain categorical labels, we also included three datasets that contain text-based descriptions of the images: EYENET (Huang et al., 2021a), ODIR-5K, and STARE (Hoover, 2000; Hoover and Goldbaum, 2003). For the datasets that contain pixel-level annotations of fundus images, these are converted to image-level labels. The resulting dataset $\mathcal{D}_T$ is thus composed of 286,916 images, which include 96 different target categories. For further details, we refer the reader to Supplementary Materials, Section S1.

***Standarization and augmentations***. All images are resized to a canvas of size $512 \times 512$, and zero-padding is applied to rectangular-shaped images to avoid distortions. Furthermore, all images are intensity-normalized to be in the $[0, 1]$ range. During training, random image augmentations are applied using horizontal flips, random rotations of $[-5, 5]$ degrees, zoom scaling in the range $[0.9, 1.1]$, and color jitter.

### *4.2. Evaluation protocol*

The proposed foundation model is validated under two different scenarios, with regard to the target task: *domain shift* (i.e., a new dataset consisting of categories that are used for training) and *unseen categories* (i.e. novel, unseen diseases that are not used during training). For these purposes, we omitted several datasets and categories during training.

Table 1: **Assembly of retinal fundus images datasets from various open-access sources.** We have developed a vision-language universal model for fundus image understanding by compiling 37 publicly available, mostly categorical datasets. The ensuing dataset assembly consists of a total of 286,916 images corresponding to 96 different categories. Among the 37 datasets, only 3 include text-based supervision of fundus images. For more detailed information on category abbreviations, we refer the reader to Supplemental Materials, Section S1

| Datasets | #Targets | #Images | Labels | Annotations |
|---|---|---|---|---|
| 01.EYEPACS[1] | 5 | 88,702 | noDR, mildDR, modDR, sevDR, prolDR. | Categorical |
| 02.MESSIDOR2 (Decencière et al., 2014; Krause et al., 2018) | 9 | 1,748 | noDR, mildDR, modDR, sevDR, prolDR, noisy, clean, DME, noDME, hEX. | Categorical |
| 03.IDRID (Porwal et al., 2020) | 10 | 597 | MA, HE, hEX, sEX, noDR, mildDR, modDR, sevDR, prolDR, noDME, nonCSDME, DME. | Categorical |
| 04.RFMid (Pachade et al., 2021) | 46 | 3,200 | 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, RPEC, MHL, RP, CWS, CB, ODM, PRH, MNF, HR, CRAO, TD, CME, PTCR, CF, VH, MCA VS, BRAO, PLQ, HPED, CL. | Categorical |
| 05.1000x39 (Cen et al., 2021) | 39 | 1,000 | N, TSLN, LOC, mildDR, modDR, sevDR, BRVO, CRVO, G, CRAO, RD, CSR, VKH, M, ERM, MHL, MYA, HE, OA, NP, sevHR, DSE, DD, CDA, RP, BCD, PRDB, MNF, VH, F, hEX, YWSF, CWS, TV, CB, LS, noisy, noProlDR, prolDR. | Categorical |
| 06.EYENET (Huang et al., 2021a) | - | 15,708 | – | Text |
| 07.LAG (Li et al., 2019a) | 2 | 4,854 | G, noG. | Categorical |
| 08.ODIR-5K [2] | ≥7 | 8,000 | N, DR, G, CAT, ARMD, HR, MYA. | Text |
| 09.PAPILA (Kovalyk et al., 2022) | 2 | 488 | G, N. | Categorical |
| 10.PARAGUAY (Castillo Benítez et al., 2021) | 7 | 1,437 | noDR, mildDR, modDR, sevDR, prolDR. | Categorical |
| 11.STARE (Hoover, 2000; Hoover and Goldbaum, 2003) | - | 397 | – | Text |
| 12.ARIA (Farnell et al., 2008) | 3 | 143 | N, ARMD, DR. | Categorical |
| 13.FIVES (Jin et al., 2022) | 6 | 800 | noisy, clean, ARMD, DR, G, N. | Categorical |
| 14.AGAR300 (Derwin et al., 2020) | 2 | 28 | DR, MA. | Categorical |
| 15.APTOS[3] | 5 | 5,590 | noDR, mildDR, modDR, sevDR, prolDR. | Categorical |
| 16.FUND-OCT (Hassan et al., 2021, 2019) | 7 | 200 | G, N, CME, neovARMD, geoARMD, acCSR, chCSR. | Categorical |
| 17.DiaRetDB1 (Kauppi et al., 2007) | 9 | 89 | IrMA, neoV, ReSD, hEX, HE, sEX, MA. | Categorical |
| 18.DRIONS-DB (Carmona et al., 2008) | 1 | 110 | noCAT, Dis. | Categorical |
| 19.Drishti-GS1 (Sivaswamy et al., 2014) | 2 | 100 | N, G. | Categorical |
| 20.E-ophta (Decencière et al., 2013) | 2 | 463 | EX, MA., | Categorical |
| 21.G1020 (Bajwa et al., 2020) | 2 | 1,020 | G, N. | Categorical |
| 22.HEI-MED (Giancardo et al., 2012) | 3 | 169 | EX, CWS, DN. | Categorical |
| 23.HRF (Budai et al., 2013) | 4 | 81 | N, G, DR, noisy. | Categorical |
| 24.ORIGA (Zhang et al., 2010) | 2 | 650 | G, noG. | Categorical |
| 25.REFUGE (Orlando et al., 2019; Li et al., 2020) | 2 | 1200 | G, noG. | Categorical |
| 26.ROC (Niemeijer et al., 2010) | 1 | 100 | MA. | Categorical |
| 27.BRSET (Nakayama et al., 2023; Goldberger et al., 2000) | 24 | 16,266 | noDR, mildDR, modDR, sevDR, prolDR, HE, hEX, sEX, MA, AOD, AV, AM, noisy, clean, ME, S, NE, ARMD, BRVO, HR, DN, HE, RD, MYA, ICD. | Categorical |
| 28.OIA-DDR (Li et al., 2019b) | 9 | 13,673 | noDR, mildDR, modDR, sevDR, prolDR, HE, hEX, sEX, MA. | Categorical |
| 29.AIROGS (de Vente et al., 2023) | 2 | 101,442 | G, noG | Categorical |
| 29.SYSU (Lin et al., 2020) | 8 | 1,220 | noDR, mildDR, modDR, sevDR, prolDR, HE, hEX, sEX. | Categorical |
| 31.JICHI (Takahashi et al., 2017) | 5 | 9,940 | noDR, mildDR, modDR, sevDR, prolDR | Categorical |
| 32.CHAKSU (Kumar et al., 2023) | 2 | 1,345 | G, noG | Categorical |
| 33.DR1-2 (Pires et al., 2014) | 7 | 1,597 | N, ReSD, hEX, DN, CWS, supHE, deepHE | Categorical |
| 34.Cataract[4] | 4 | 601 | N, G, CAT, RS | Categorical |
| 35.ScarDat (Wei et al., 2018) | 2 | 997 | LS, noLS | Categorical |
| 36.ACRIMA (Diaz-Pinto et al., 2019) | 2 | 705 | G, noG | Categorical |
| 37.DeepDRiD (Liu et al., 2022) | 5 | 2,256 | noDR, mildDR, modDR, sevDR, prolDR | Categorical |
| | ≥96 | 286,916 | | |

***Domain shift***. We used three datasets, consisting of the main fundus image analysis tasks, for testing. Specifically, these are removed from the training data, and are: The **MESSIDOR** dataset for the task of diabetic retinopathy grading, the **REFUGE** dataset for glaucoma detection, and the **FIVES** dataset for the classification of heterogeneous diseases.

***Unseen categories***. To evaluate the zero-shot generalization and transferability of our method to novel diseases, we selected four categories and removed them from the training data: retinitis pigmentosa (RP), macular hole (MHL), cataract (CAT) and pathologic myopia (MYA) (see Supplementary Material, Figure S2, for the visualization of these conditions). Thus, two subsets are created: **20x3**, which contains 20 samples for normal, RP, and MHL retrieved from the 1000x39 dataset, and **ODIR200x3**, which contains 200 images for normal, CAT, and MYA retrieved from the ODIR-5K dataset. It is worth mentioning that any sample corresponding to the novel categories was not used during the training of the foundation model. Table 2 provides a summary of the main datasets used for evaluation.

***Generalization after adaptation***. We used several additional datasets to evaluate the generalization performance of the foundation model after tuning. In particular, we used **ACRIMA** for glaucoma detection, **DeepDRiD** for diabetic retinopathy grading, and the **RP-MHL-2** and **CAT-MYA-2** subsets for the novel categories. A detailed description of these additional subsets is presented in Supplementary Materials, Section S1 and Table S1.

Table 2: **Dataset distribution for evaluating the foundation model**. To evaluate the generalization capabilities of FLAIR, we removed several datasets from the training phase. The evaluation scenarios are: *i)* domain (image-data) shifts on classes that are seen during training, and *ii)* novel, unseen categories. For the latter scenario, the samples corresponding to the new target categories (i.e., RP, MHL, CAT and MYA) were not used during training.

| Dataset | #Images | Labels |
|---|---|---|
| *Domain shift* | | |
| MESSIDOR | 1448 | noDR, mildDR, modDR, sevDR, prolDR. |
| FIVES | 800 | N, DR, G, ARMD. |
| REFUGE | 1200 | G, noG |
| *Unseen categories* | | |
| 20x3 | 60 | N, RP, MHL |
| ODIR200x3 | 600 | N, CAT, MYA |

### *4.3. Foundation model pre-training*

We designed the vision and language encoders following previous relevant literature on vision-language pre-training for medical images. In particular, we used ResNet-50 (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009) as a vision encoder, $\theta_f$, following ConVIRT (Zhang et al., 2022b), GloRIA (Huang et al., 2021b) and MedCLIP (Wang et al., 2022b). The text encoder $\phi_f$ is BioClinicalBERT[5], similarly to MedCLIP (Wang et al., 2022b). For both encoders, linear layers with 512 output features are used as projection heads, $\theta_p$ and $\phi_p$. The proposed vision-language pre-training is performed on the assembly $\mathcal{D}_T$ dataset by minimizing Eq. 3 during 15 epochs, using a batch size of 128 images. We use AdamW with a base learning rate of $1e^{-4}$, and a warm-up cosine scheduler during the first epoch. We employ domain knowledge descriptors for categorically-labeled data to map labels into text. For more details on the used descriptors, please refer to Table S3. Hereafter, we refer to the proposed foundation model as FLAIR-$\pi_{EK}$. The training is carried out using mixed precision, on a single RTX A6000 card, and it takes about 16 hours.

### *4.4. Baselines*

In the following, we describe the different baselines and methods used to assess the performance of the proposed foundation model. In particular, we benchmark the generalization and transferability capabilities of FLAIR-$\pi_{EK}$ against other relevant strategies. Concretely, we compare to *i)* vision-language models for zero-shot generalization, *ii)* vision encoder pre-training for efficient transfer learning, and *iii)* fully model training for each target dataset, instead of following an efficient adaptation strategy.

#### *4.4.1. Language-driven zero-shot classification*

***Vision-language pre-training (VLP)***. First, we define the baselines for the task of language-driven (i.e., zero-shot) classification. We use CLIP (Radford et al., 2021) with its original weights, pre-trained on 400M image-text pairs from heterogeneous sources. It is worth mentioning that CLIP pre-training included several medical imaging datasets. In addition, we include BiomedCLIP (Zhang et al., 2023), a recently published generalist vision-language model for understanding biomedical images. This model follows a strategy that is currently gaining increasing popularity, in which the vision-language pre-training is carried out with large biomedical imaging datasets from highly diverse sets of modalities and tasks, e.g. radiographs, CT volumes, histology, dermatology, ultrasound images, etc., in order to obtain a common foundation model for all biomedical fields. Thus, BiomedCLIP is pre-trained using more than 15 million image-text pairs from heterogeneous medical domains, extracted from PubMed. Also, we use a basic version of the proposed foundation model, which does not integrate domain-knowledge descriptors in the training. We refer to this as FLAIR-$\pi_{naive}$. Note that, for this basic version, the training implementations are the same as those used for the proposed knowledge-driven foundation model.

#### *4.4.2. Pre-train-and-adapt baselines*

***Task-specific models (TSMs)***. To evaluate the benefits of incorporating language supervision during pre-training across multiple tasks, we set as baselines task-specific models, trained from the assembly of datasets. A different model is trained for each of the main evaluated tasks: diabetic retinopathy grading ($\text{TSM}_{DR}$ model), glaucoma detection ($\text{TSM}_{Glaucoma}$ model), and multiple disease classification ($\text{TSM}_{Diseases}$ model) - see Table 2 for additional details on each task. For each task, all samples categorically labeled with the corresponding task-target categories are retrieved from the assembly dataset. For example, $\text{TSM}_{DR}$ model is pre-trained using all images labeled as noDR, mildDR, modDR, sevDR, and prolDR, from the dataset assembly. The resulting pre-training sub-datasets, per task, are composed of nearly $\sim 100,000$ images for $\text{TSM}_{DR}$ and $\text{TSM}_{Glaucoma}$ tasks and nearly $\sim 10,000$ samples for $\text{TSM}_{Diseases}$. The task-specific models are trained using a standard multi-class cross-entropy loss.

***Other pre-training baselines***. Finally, we examined additional baselines for learning pre-trained representations. Concretely, we used the features extracted from pre-trained ResNet-50 on ImageNet (He et al., 2016) for adaptation, which we refer to as *ImageNet*. In addition, we evaluated contrastive unsupervised learning using SimCLR (Chen et al., 2020) as a pre-training strategy in our scenario. Here, SimCLR is trained on the whole assembly dataset, using a batch size of 64 images, during 15 epochs. The same transformations used for our foundation model are applied in this setting for positive-pair augmentations.

#### *4.4.3. Fully-training upper-bound*

***Dataset-specific models (Supervised)***. As an upper bound, we train dataset-specific models on the target datasets. This

[1] https://www.kaggle.com/c/diabetic-retinopathy-detection
[2] https://odir2019.grand-challenge.org/
[3] https://www.kaggle.com/c/aptos2019-blindness-detection
[4] https://www.kaggle.com/datasets/jr2ngb/cataractdataset
[5] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT

means, instead of zero-shot generalization or lightweight adaptation of a pre-trained model, all model weights are tuned on the target dataset. More concretely, the backbone initialization is the same as the foundation model, using ResNet-50 with initialization weights pre-trained on ImageNet (He et al., 2016), as it is a popular approach for transfer learning to medical images (Raghu et al., 2019). A batch size of 8 images is used, and the model is trained via standard stochastic gradient descent during 50 epochs using a linear classifier and multi-class cross-entropy loss. Other hyperparameters are ADAM optimizer and a learning rate of $1e^{-4}$. We refer to this strategy as *Supervised*.

### *4.5. Evaluation metrics*

The main figure of merit used for evaluation is accuracy per class, and the balanced average for each dataset (ACA) (Zhao et al., 2019). In addition, task-specific metrics are incorporated into the evaluation for comparison with previous literature. In particular, DR grading is evaluated using the quadratic Cohen kappa, the most popular choice in previous relevant literature (Galdran et al., 2020). In the case of Glaucoma detection, the area under receiving-operative-curve (AUC) is used as a figure of merit, following previous challenges on this topic (Orlando et al., 2019; de Vente et al., 2023). For all the transferability experiments that require training or adaptation, metrics are averaged across 5 cross-validation folds.

## 5. Results

### *5.1. Generalization*

In this section, we evaluate the generalization capabilities of the proposed model by direct prediction (i.e. no adaptation of the trainable parameters) on several target datasets under two common scenarios: *i)* domain shift, where the set of classes remains the same, but images present domain drifts, and *ii)* novel classes, where the domain remains the same, but unseen classes are expected to be identified.

#### *5.1.1. Presence of domain shift*

First, we assess the performance of the proposed pre-trained vision-language approach, FLAIR, under domain distributional shifts, which is benchmarked against task-specific models. To achieve this, we consider two inference possibilities: using a naive mapping prompt consisting of the category names ($\pi_{naive}(\cdot)$), and domain-knowledge descriptions of the target diseases ($\pi_{EK}(\cdot)$). These scenarios are referred to as *VLP-inference w/*$\pi_{naive}$ and *VLP-inference w/*$\pi_{EK}$, respectively. Furthermore, to further understand the impact of integrating domain-knowledge descriptions during training, we evaluate our model when both naive and the proposed mapping are used, whose models are referred to as FLAIR-$\pi_{naive}$ and FLAIR-$\pi_{EK}$.

The results from this experiment are presented in Table 3. First, we can observe that standard vision-language pre-training (VLP), i.e., FLAIR-$\pi_{naive}$, provides comparable results to those obtained by traditional supervision models on DR and Glaucoma tasks. In contrast, vision-language pre-training methods outperform by a large margin (+34%) its standard supervised counterpart, $\text{TSM}_{Diseases}$, on the FIVES dataset. It is important to note that the task-specific models $\text{TSM}_{DR}$ and $\text{TSM}_{Glaucoma}$ were trained with a large number of labeled target task samples, whereas the size of the available dataset for $\text{TSM}_{Diseases}$ was much smaller. This may explain the significant discrepancies in performance differences between task-specific and VLP models across tasks. Furthermore, these results showcase a clear benefit of pre-trained vision-language models, as more datasets covering a wide variability can be used during training, circumventing the problem of small labeled datasets for a specific task. The second important observation from Table 3 is that introducing domain-knowledge descriptions in the proposed foundation model (FLAIR-$\pi_{EK}$) typically yields significant improvements on MESSIDOR (diabetic retinopathy grading) and REFUGE (glaucoma detection), whereas the performance on FIVES (diseases classification) is slightly degraded. We advocate that the large improvement gains observed are due to introducing hierarchical relationships between local findings in DR grades via integrating the domain-knowledge prompts in FLAIR-$\pi_{EK}$. Finally, the proposed approach obtains results that are on par with task-specific solutions in the literature, such as cost-sensitive optimization for DR grading (Galdran et al., 2020) and optic-disk segmentation for Glaucoma detection (AIROGS leaderboard (de Vente et al., 2023)). In con-

Table 3: **Generalization under domain shifts.** The results obtained by direct prediction, i.e., no adaptation, of the pre-trained language-driven models under the presence of domain shifts. The proposed approach, FLAIR, is compared to task-specific models (trained following the standard fully-supervised paradigm), and the existing literature on the corresponding specific tasks: diabetic retinopathy (MESSIDOR), disease classification (FIVES) and glaucoma detection (REFUGE). For each task, we provide representative figures of merit in the literature. The proposed method, FLAIR-$\pi - EK$ is shadowed, whereas the best results are highlighted in bold.

| Method | Dataset | | |
|---|---|---|---|
| | MESSIDOR | FIVES | REFUGE |
| | $^{DR\ grading}$ (ACA/$\kappa$) | $^{Diseases}$ (ACA) | $^{Glaucoma}$ (AUC) |
| *Prior literature* | | | |
| $\text{DR}_{graduate}$ (Araújo et al., 2020) | 0.596/0.710 | - | - |
| AST (Galdran et al., 2020) | 0.634/0.797 | - | - |
| $\text{AIROGS}_{lb}$ (de Vente et al., 2023) | - | - | [0.88, 0.94] |
| *Task-specific models (TSMs)* | | | |
| $\text{TSM}_{DR}$ | 0.550/**0.772** | - | - |
| $\text{TSM}_{Diseases}$ | - | 0.381 | - |
| $\text{TSM}_{Glaucoma}$ | - | - | 0.904 |
| *VLP - inference w/* $\pi_{naive}$ | | | |
| CLIP | 0.237/0.140 | 0.250 | 0.470 |
| BiomedCLIP | 0.224/0.201 | 0.416 | 0.540 |
| FLAIR-$\pi_{naive}$ | 0.545/0.662 | 0.732 | 0.899 |
| FLAIR-$\pi_{EK}$ | 0.602/0.711 | 0.719 | 0.918 |
| *VLP - inference w/* $\pi_{EK}$ | | | |
| CLIP | 0.200/0.000 | 0.256 | 0.433 |
| BiomedCLIP | 0.207/0.188 | 0.415 | 0.624 |
| FLAIR-$\pi_{naive}$ | 0.442/0.694 | **0.744** | 0.871 |
| FLAIR-$\pi_{EK}$ | **0.604**/**0.772** | 0.735 | **0.920** |

trast, the proposed model is universal, providing a general representation of fundus images, which results in important performance gains even compared to standard supervised methods in targeted-task scenarios. We stress that, due to specific implementation differences, it might be difficult to establish direct comparisons with prior works in the standard targeted-task setting.

### *5.1.2. Performance on novel classes*

We now present empirical evaluations of FLAIR in the zero-shot scenario, i.e. there is no adaptation of the foundation model to novel, unseen categories. In this context, we study three different strategies for text-prompt design. First, we present the prompt generation as an anomaly detection task, in which all unseen diseases are treated as anomalies. In this case, the prompts used as input of the text encoder are simply either "*normal*" or "*disease*". Then, we introduce the notion of each novel disease when generating the text prompts. The first strategy involves using directly the names of the unseen diseases as prompts, in a *naive* way. Last, and to further exploit the learning power of vision-language models, we propose to *design* domain knowledge prompts, which briefly describe the differential findings on fundus images corresponding to each condition. For example, instead of employing the category name *"cataract"*, as in the naive manner, we use the following finding as input text prompt: *"opacities in the macular area"*. In Supplementary Material, Fig. S2, we provide additional prompts designed for the unseen categories.

Table 4: **Generalization to unseen categories (zero-shot classification).** Transferability of the proposed foundation model to new tasks via prompt-based inference. We evaluated three different strategies for generating text prompts: anomaly detection (i.e., "*normal*"/"*disease*"), classification via naive prompt (i.e., the new disease name) and designed prompts (i.e., domain-knowledge descriptors). The metric presented is the accuracy for each category. The proposed method, FLAIR-$\pi_{EK}$, is shadowed, whereas the best results are highlighted in bold.

| Method | Dataset | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 20x3 | | | | ODIR200x3 | | | |
| | N | RP | MHL | Avg. | N | CAT | MYA | Avg. |
| *Anomaly Detection Inference (i.e. "normal/disease")* | | | | | | | | |
| CLIP | 1.000 | 0.200 | | 0.600 | 0.770 | | 0.412 | 0.591 |
| BiomedCLIP | 0.950 | 0.125 | | 0.538 | 0.800 | | 0.770 | **0.785** |
| FLAIR-$\pi_{naive}$ | 0.900 | 0.200 | | 0.550 | 1.000 | | 0.102 | 0.551 |
| FLAIR-$\pi_{EK}$ | 0.850 | 0.775 | | **0.812** | 0.985 | | 0.350 | 0.668 |
| *Inference with Naive Prompts - $\pi_{naive}$ (e.g. "cataract")* | | | | | | | | |
| CLIP | 0.100 | 1.000 | 0.000 | 0.367 | 0.770 | 0.495 | 0.070 | 0.445 |
| BiomedCLIP | 0.900 | 0.950 | 0.400 | 0.750 | 0.765 | 0.920 | 0.495 | **0.727** |
| FLAIR-$\pi_{naive}$ | 0.950 | 0.650 | 0.100 | 0.567 | 0.990 | 0.340 | 0.010 | 0.447 |
| FLAIR-$\pi_{EK}$ | 0.950 | 0.600 | 1.000 | **0.850** | 0.990 | 0.455 | 0.005 | 0.483 |
| *Inference with Expert Knowledge Prompts - $\pi_{EK}$ (e.g. "opacity in the macular area")* | | | | | | | | |
| CLIP | 1.000 | 0.000 | 0.000 | 0.333 | 0.290 | 0.195 | 0.955 | 0.480 |
| BiomedCLIP | 0.400 | 0.800 | 0.650 | 0.617 | 0.125 | 0.695 | 0.930 | 0.583 |
| FLAIR-$\pi_{naive}$ | 1.000 | 0.900 | 0.050 | 0.650 | 0.405 | 0.015 | 0.990 | 0.470 |
| FLAIR-$\pi_{EK}$ | 1.000 | 0.950 | 1.000 | **0.983** | 0.760 | 0.765 | 0.475 | **0.667** |

Table 4 reports the results. One may observe that training CLIP on the assembly dataset with categorical prompts (i.e., FLAIR-$\pi_{naive}$) typically yields large performance gains in comparison to the standard CLIP model, particularly in the multi-class classification scenario (*middle and bottom sections of the Table*). However, FLAIR-$\pi_{naive}$ largely fails to differentiate between the novel categories, acting as a normal fundus image detector. In contrast, integrating domain-specific knowledge brings substantial performance gains. First, using specific domain descriptors during training results in +29% and +4% of improvement over the naive FLAIR version (see Table 4, *Prompt naive* section). The observed large improvements may be explained by the fact that the text prompts used in the proposed $\pi_{EK}$ favor a richer text embedding, with hierarchical and domain-specific knowledge. Interestingly, further leveraging these better-designed prompts during inference enhances the different models by around 13% (20x3) and 18% (ODIR200x3), resulting in a gap with FLAIR-$\pi_{naive}$ of nearly 33% in 20x3 dataset, and 10% in ODIR200x3 (see Table 4, section *Expert Knowledge Prompts*). More significantly for the medical domain, the proposed model substantially outperforms CLIP across all the scenarios, with differences going up to 65%. This shows that universal vision-language models trained on general computer vision tasks yield suboptimal results in specialized medical imaging fields. The obtained results demonstrate that, in the absence of large datasets with text-based supervision, samples with categorical labels could still be exploited to train powerful vision-language representations, by encoding domain expert knowledge into text supervision. Furthermore, FLAIR reaches promising performances on the anomaly detection task, which does not require defining the diseases on the target dataset. Finally, it is worth mentioning the limitations observed using a generalist model, such as BiomedCLIP, for zero-shot generalization on retinal fundus images. First, even though this model improves CLIP consistently across all the tasks (see Tables 3 and 4), the results obtained in comparison to the specialized model (i.e. FLAIR-$\pi_{EK}$) are considerably lower, especially in the tasks reported in Table 3. Furthermore, while BiomedCLIP presents a competitive performance on one dataset for generalizing to unseen categories (i.e., ODIR200x3), its results degrade when using descriptive prompts at inference, which highlights its limited ability to encode specialized hierarchical expert knowledge. These results showcase that, despite the focus of the recent literature on generalist models, the use of domain-specialized models such as FLAIR shows more promising results in the context of fundus imaging. We anticipate that this might be case in other medical-imaging domains. It should be noted, however, that it is difficult to establish comparisons with BiomedCLIP, since the lack of transparency in the description of these massive databases makes it difficult to know whether the model has been trained on tasks used for testing or not.

## *5.2. Transferability*

We now evaluate the capabilities of FLAIR-$\pi_{EK}$ to transfer the learned representations to downstream domains and tasks using minimal, efficient fine-tuning. Thus, we cover the scenario in which the trained feature extractor is frozen during adaptation, and just a lightweight module with trainable parameters (i.e. adapter) is added on top. More concretely, two strategies are evaluated in this scenario: *i)* Linear Probing (Radford et al., 2021) using only the image modality, and

*ii)* vision-language adapters, which combine both modalities during adaptation.

### 5.2.1. Linear Probing

***Adaptation strategy***. We use the features extracted from the vision encoder $\theta_f(\cdot)$ as input of an additional linear classifier, whose parameters are fine-tuned during adaptation. This strategy, which is commonly referred to as Linear Probe (LP) in the literature, is further validated empirically in the ablation experiments presented in Section 5.2.3.

***Adaptation resources***. To evaluate common scenarios, two data regimes are studied: *i) low data regime*, in which only a few support (i.e., labeled) images for each category in the target dataset are retrieved for adaptation ($k = \{1, 5, 10\}$), and *ii)*: *large data regime*, where a large percentage of the dataset is used during adaptation, i.e. $\{20\%, 40\%, 60\%, 80\%\}$.

***Training/testing splits***. The test subset remains the same across all data regimes, using 20% of the target dataset. From the training subset, only a number of samples corresponding to the target data regime are randomly retrieved for adapting the pre-trained model. This process is averaged across 5 cross-validation folds.

***Baselines***. To validate the benefits of the proposed foundation model, other common transfer learning strategies are evaluated: *i)*: LP of task-specific models (TSMs), in which the model is pre-trained using only all samples that contain a subset of classes included in the assembly dataset, in a standard supervised way, *ii)*: LP with unsupervised pre-training, using SimCLR, *iii)*: LP with transfer learning from natural images, using the backbone pre-trained on ImageNet, and *iv)*: dataset-specific models via standard supervised training (i.e. full network tuning in the target dataset) with weights pre-trained on ImageNet. For more details on the different baselines, we refer the reader to Sections 4.4.2 and 4.4.3.

The results for transferability through image features adaptation on the evaluation datasets are presented in Figure 4.

***Task-specific models (TSMs)***. The obtained results unveil the limitations of task-specific models regarding transferability. While they perform well for the task they have been trained on, with enough pre-training data (see Figure 4: $\text{TSM}_{DR}$ on MESSIDOR or $\text{TSM}_{Glaucoma}$ on REFUGE), their performance degrades in more challenging scenarios. Concretely, this occurs when *i)*: the pre-training dataset is relatively smaller (see Figure 4: $\text{TSM}_{Diseases}$ on FIVES), and *ii)*: the models are adapted to unseen tasks (see Figure 4: $\text{TSM}_{Glaucoma}$ on MESSIDOR, or $\text{TSM}_{DR}$ on 20x3).

***Dataset-specific models (supervised)***. Also, from Figure 4, one may observe the limitations of training dataset-specific models. This strategy performs poorly in the low-data regime (see Figure 4: all datasets). Even in the larger data regime, the available samples might not be enough to reach the performance obtained by FLAIR (see Figure 4: Supervised on MESSIDOR). In addition, they are computationally expensive, since they require tuning the whole model.

***Proposed foundation model (FLAIR)***. The main takeaways on the performance of LP adaptation using FLAIR pre-training on the assembly dataset are: *i)* It performs on par with the best task-specific models under domain shift (see Figure 4: MESSIDOR dataset); *ii)* It outperforms by a large mar-

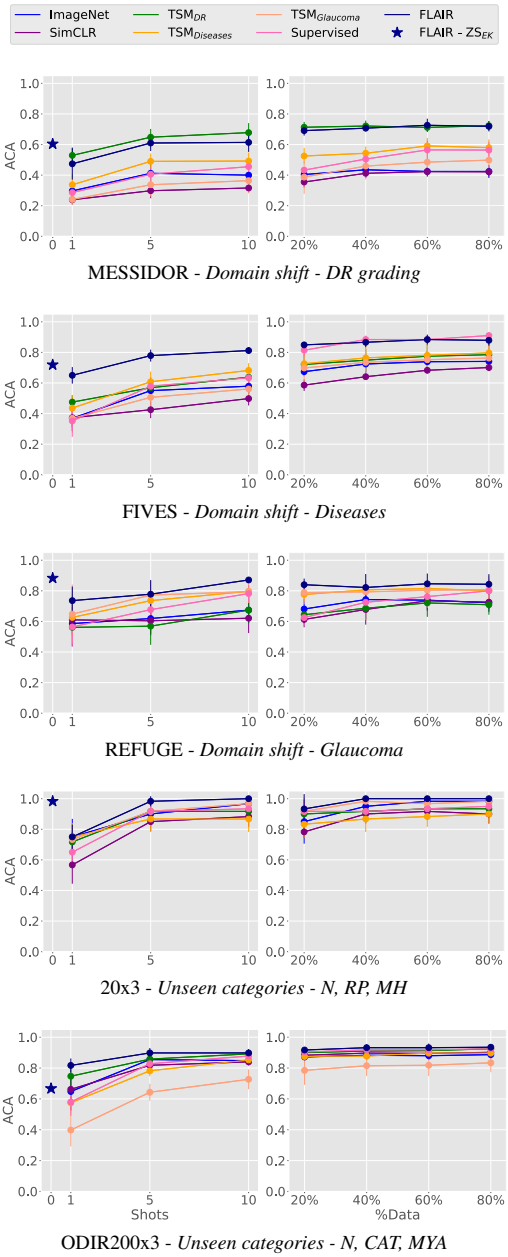



Figure 4: **Transferability.** The results of transferring the feature representations of the pre-trained models to downstream domains and tasks in the low-data (*left column*) and large-data (*right column*) regimes. The results were obtained by adjusting a linear-probe classifier. The metric presented is the average accuracy, averaged across 5 cross-validation folds. ZS: zero-shot (i.e., prompt-based classification).

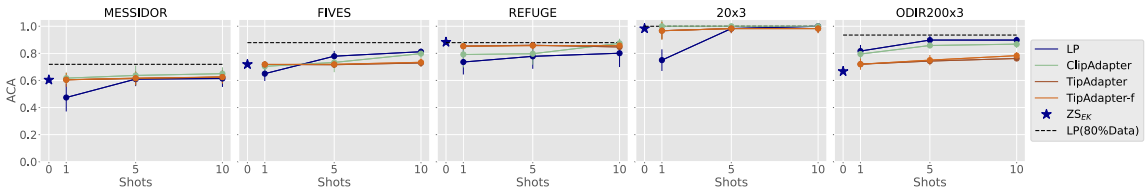



Figure 5: **Vision-language few-shot adapters.** The results of different adapters in the few-shot setting. The metric presented is the average accuracy, averaged across 5 cross-validation folds. ZS: zero-shot (i.e., prompt-based classification with domain-knowledge prompts).

gin these models on tasks under-represented in the assembly dataset (see Figure 4: FIVES dataset); *iii)* On unseen categories, it outperforms adaptation of the supervised dataset-specific models in the low data regime (see Figure 4: 20x3 and ODIR200x3 datasets); and *iv)*. In many cases, it also outperforms these fully-tuned models in the large data regime (see Figure 4: MESSIDOR, REFUGE and 20x3 datasets), which alings with relevant recent literature on vision-language pre-training for radiology imaging (Tiu et al., 2022).

### 5.2.2. *Vision-language adapters*

Recent emergent literature in computer vision has investigated strategies, often referred to as adapters, to fine-tune vision-language models in low-data (few-shot) regimes for the target tasks, e.g., Clip-Adapter (Gao et al., 2021) and Tip-Adapter (Zhang et al., 2022a). These strategies typically integrate the knowledge driven from the pre-trained language encoder along with the vision features and use additional layers in the networks. Still, the utility of these adapters remains largely unexplored in the medical domain. Figure 5 depicts the results obtained by different vision-language adapters using our pre-trained FLAIR foundation model and expert-knowledge prompts, across the different tasks. The results point to the powerful capabilities of zero-shot classification in different scenarios. In most of the cases, zero-shot inference, enhanced with domain-expert knowledge prompts, outperforms adaptation using $k \leq 5$ shots (see Figure 5 MESSIDOR, FIVES, REFUGE, 20x3). As for the vision-language adapters (Zhang et al., 2022a; Gao et al., 2021), these do not seem to provide consistent improvements, neither over zero-shot classification (when $k \leq 5$) nor over basic Linear Probing (when $k = 10$).

### 5.2.3. *Ablation experiments*

In this section, we present ablation experiments that motivate different decisions in the design of the proposed framework.

***What features to use for knowledge transfer***. Vision-language pre-training models align the image-encoder features, $\theta_f(\cdot)$, to the text representations via a projection, $\theta_p(\cdot)$, along with a mapping to the unit hyper-sphere using an l2 normalization. Regarding the transferability of the pre-trained visual features to downstream domains and tasks via linear probing (LP), the standard feature-representation choice in prior literature is often based on both projection and normalization (Radford et al., 2021; Gao et al., 2021; Zhang et al., 2022a). In the following ablation experiment, we evaluate the feature transferability for the different evaluation datasets using the following three options: vision, projected, and projected-and-normalized features. We evaluated the three options under both the low and large-data regimes, using $k = 10$ and 80% of the dataset for training.

Figure 6 depicts the results, which show performance improvements across most of the tasks when using visual representation $\theta_f(\cdot)$ for transferability, in comparison to using projected features $\theta_p(\cdot)$ or projected-and-normalized features $\theta_p(\cdot)$ + norm. Motivated by these observations, we selected original feature representation $\theta_f(\cdot)$ for the transferability experiments in this work.

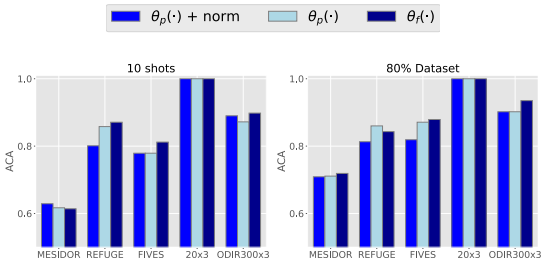



Figure 6: **Ablation experiment on the transferred features for adaptation.** Evaluation of the performance of the linear-probe transferability of the features extracted from the vision encoder, $\theta_f(\cdot)$, the inter-modality projection head, $\theta_p(\cdot)$, and its hypersphere normalization, $\theta_p(\cdot)$ + norm. The metric presented is the average accuracy, averaged across 5 cross-validation folds. The results are presented for the low-data (10 shots) and large-data (80% of the whole dataset) regimes.

***Generalization of linear-probe adaptation under domain shifts***. The *pre-train-and-adapt* strategy using image-language models and computationally efficient linear-probe adaptation has shown promising performances on downstream computer-vision tasks. In the following, we aim at conducting a more comprehensive evaluation of this linear-probe strategy, in order to assess the capacity of the adaption stage in response to new changes in a target domain (i.e., there are domain shifts after adaptation). To do so, we employ the supplementary evaluation subsets; see Table S1. In particular, we evaluate the performance of the linear probe, which has

been fine-tuned on a source domain, in a novel target domain. More concretely, the adaptation is performed as follows using two datasets A and B: The model is fine-tuned on A and tested on B, and vice-versa. Again, two feature representations are evaluated for transferability: features extracted from the vision encoder, $\theta_f(\cdot)$, and features based on the inter-modality projection head, $\theta_p(\cdot)$. We juxtapose the performance of the linear probe to fine-tuning all the model trainable parameters on the source data (i.e., using a standard supervised-learning setting, but with parameter initialization using either FLAIR or Imagenet model), as well as to the zero-shot performance. The experiments are carried out in the large-data regime, to evaluate the best-case scenario, in which the available data is not a limiting factor.

Figure 7 depicts the results from these experiments, which point to the following takeaways: The supervised, dataset-specific models, which update all the trainable parameters using the source data, reach good performance on the source domain. However, they struggle to generalize under domain shifts; see Figure 7, *first*, and *second rows*. Linear probing (LP) from the foundation model mitigates this difficulty in several cases; see Figure 7, *first, second*, and *third rows*. These observations emphasize the generalization capabilities of the foundation model and are in line with recent observations in the computer vision community (Kumar et al., 2022). This is especially the case when adapting the feature representation of the multi-modal projection, $\theta_p(\cdot)$. As echoed earlier in Figure 6, the vision-encoder features, $\theta_f(\cdot)$, seem more specialized for the source domain and, hence, yielded the best overall performance. However, the experimental results in Figure 7 suggest that the performances yielded by these features could be affected under domain shifts; see Figure 7, *all*. Thus, what representation to use for adaptation might depend on the expected data variability and the consistency between the source and target domains over time. It is worth mentioning that, when a large domain shift is expected, the prompt-based (zero-shot) classification might result in a robust solution. In many scenarios, the performance of LP on a target domain is below the prompt-based classification. In the case of known categories by the foundation model, prompt-driven classification in a zero-shot fashion achieves even more promising results for different domains; see Figure 7, *first* and *second rows*. We provide additional results in Section S2.2, which illustrate the robustness of text-driven transferability to domain shifts and are in line with recent observations in the computer-vision literature (Wortsman et al., 2022; Goyal et al., 2023). Of course, and as echoed in Figure 5, this robustness of zero-shot classification to domain variability comes at the price of a source-domain performance that is lower than LP.

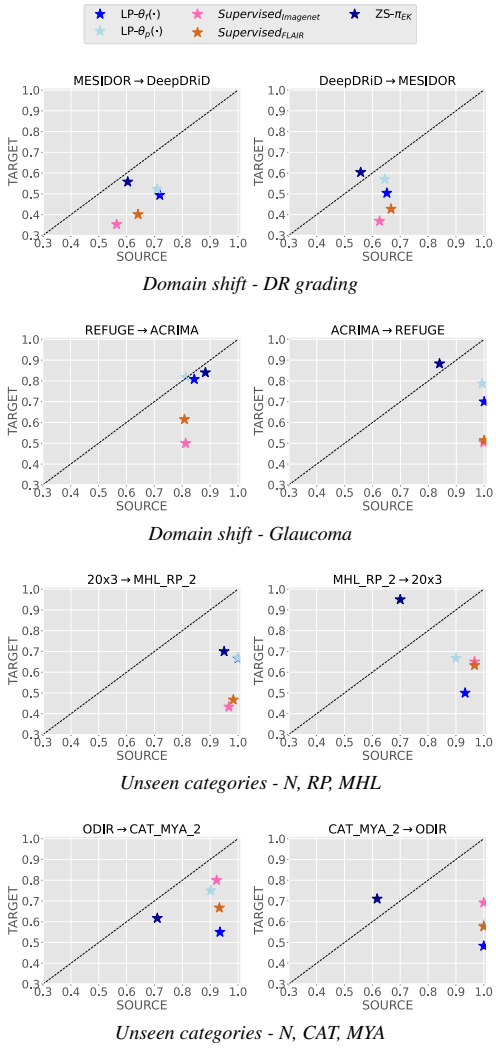



Figure 7: **Generalization of the transferred features after the adaptation stage** Evaluation of the linear-probe generalization with respect to the transferred features. Adapters are tuned on the source domain, and the performance is evaluated on another dataset with the same categories. The metric presented is the average accuracy, averaged across 5 cross-validation folds. The results were obtained under the large-data regime, with 80% of the training data.

## 6. Discussion

We introduced FLAIR, a novel vision-language foundation model for universal pathology detection and classification in retinal fundus images. Encoding expert's domain knowledge in the form of text-prompt supervision, FLAIR is trained on an assembly of 37 publicly available, mostly categorical datasets, containing up to 96 different target categories. By leveraging domain knowledge, we mitigated the scarcity of text-based supervision in retinal fundus imaging datasets, opening a promising avenue towards large-scale vision-language pre-training in this application domain. Specifically, we enhanced the categorical information in the datasets by text-based encoding of the major features of the pathologies as well as the hierarchies and relationships between them. Such valuable expert knowledge could be extracted from the relevant clinical literature and community standards.

We have empirically evaluated FLAIR capabilities for generalization and transferability, under scenarios with domain shifts and for unseen diseases. The proposed model shows a strong generalization, driven by domain knowledge prompts for zero-shot prediction. This is especially the case when tested on novel pathologies, in which basic categorical text prompts are uninformative. Additionally, with a lightweight, linear-probe adaptation, FLAIR outperforms fully fine-tuned, dataset-specific models on the target domains and tasks. The difference is even more pronounced under the low-data (few-shot) regime.

We also conducted comprehensive ablation studies, which show the substantial effect of integrating domain knowledge during both vision-language pre-training and zero-shot predictions. Our FLAIR model, along with domain-knowledge prompts for zero-shot prediction, outperformed significantly Contrastive Language-Image Pre-training (CLIP) on general computer vision data. Furthermore, it bypasses by margins a version of CLIP trained on the same retinal-imaging data but with naive categorical text prompts. Our results point to the potential of embedding domain-specific, expert knowledge in building vision-language foundation models targeted at different sub-domains of medical imaging, even beyond the fundus application domain tackled in this study. We showed that, even in the absence of large datasets with text-based supervision, samples with categorical labels could still be exploited to train powerful vision-language representations, by encoding expert's domain knowledge into text supervision.

Finally, we have provided in-depth experiments to deepen our understanding of the potential and limitations of the proposed methodology. Thus, we have evaluated the zero-shot and adaptation performance on supplementary datasets, through different ablation experiments. This has revealed certain limitations of the current vision-language pre-training paradigm in the medical field. First, the zero-shot generalization is sensitive to the designed prompts when tested on novel categories. While promising, the results still show certain variability, which is in line with recent observations in other works (Wang et al., 2022b). Second, although the foundation model has shown promising adaptation to new tasks with little resources, it may have difficulty transferring properly to out-of-distribution data, an observation that has also been stressed in recent vision-language literature (Wortsman et al., 2022; Goyal et al., 2023).

The encountered limitations might inspire further research directions to improve the performances of vision-language foundation models. Developing new reliable tools for language processing on expert domains, such as fundus imaging diagnosis, is an appealing future direction, which may improve the robustness of the text encoders. Finally, integrating novel adapters, able to generalize well to out-of-distribution data, might improve the performances of this emerging pre-train-and-adapt paradigm. For this challenge, text-driven adapters might be an interesting research avenue for the future.

## Acknowledgments

The work of J. Silva-Rodríguez was partially funded by the *Fonds de recherche du Québec (FRQ)* under the Post-doctoral Merit Scholarship for Foreign Students (PBEEE). The work is supported, in part, by PROMPT Quebec, via its PARTNERSHIP-AI program. We also thank Calcul Quebec and Compute Canada.

## References

Abramoff, M.D., Lou, Y., Erginay, A., Clarida, W., Amelon, R., Folk, J.C., Niemeijer, M., 2016. Improved automated detection of diabetic retinopathy on a publicly available dataset through integration of deep learning. Investigative Ophthalmology and Visual Science 57, 5200–5206.

Allen, D., Vasavada, A., 206. Cataract and surgery for cataract. British Medical Journal 333, 128–132.

Araújo, T., Aresta, G., Mendonça, L., Penas, S., Maia, C., Carneiro, A., Mendonça, A.M., Campilho, A., 2020. Dr|graduate: Uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images. Medical Image Analysis 63, 101715.

Bajwa, M.N., Singh, G.A.P., Neumeier, W., Malik, M.I., Dengel, A., Ahmed, S., 2020. G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection, in: International Joint Conference on Neural Networks (IJCNN), pp. 1–7.

Balyen, L., Peto, T., 2019. Promising artificial intelligence–machine learning–deep learning algorithms in ophthalmology. Asia-Pacific Journal of Ophthalmology 8, 264–272.

Bellemo, V., Lim, Z.W., Lim, G., Nguyen, Q.D., Xie, Y., Yip, M.Y., Hamzah, H., Ho, J., Lee, X.Q., Hsu, W., Lee, M.L., Musonda, L., Chandran, M., Chipalo-Mutati, G., Muma, M., Tan, G.S., Sivaprasad, S., Menon, G., Wong, T.Y., Ting, D.S., 2019. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in africa: a clinical validation study. The Lancet Digital Health 1, e35–e44.

Bodenreider, O., 2004. The unified medical language system (umls): Integrating biomedical terminology. Nucleic Acids Research 32.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901.

Budai, A., Bock, R., Maier, A., Hornegger, J., Michelson, G., 2013. Robust vessel segmentation in fundus images. International Journal of Biomedical Imaging , 154860.

Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vayá, M., 2019. Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical Image Analysis 66.

Carmona, E.J., Rincón, M., García-Feijoó, J., de-la Casa, J.M.M., 2008. Identification of the optic nerve head with genetic algorithms. Artificial Intelligence in Medicine 43, 243–259.

Castillo Benítez, V.E., Castro Matto, I., Mello Román, J.C., Vázquez Noguera, J.L., García-Torres, M., Ayala, J., Pinto-Roa, D.P., Gardel-Sotomayor, P.E., Facon, J., Grillo, S.A., 2021. Dataset from fundus images for the study of diabetic retinopathy. Data in Brief 36, 107068.

Cen, L.P., Ji, J., Lin, J.W., Ju, S.T., Lin, H.J., Li, T.P., Wang, Y., Yang, J.F., Liu, Y.F., Tan, S., Tan, L., Li, D., Wang, Y., Zheng, D., Xiong, Y., Wu, H., Jiang, J., Wu, Z., Huang, D., Shi, T., Chen, B., Yang, J., Zhang, X., Luo, L., Huang, C., Zhang, G., Huang, Y., Ng, T.K., Chen, H., Chen, W., Pang, C.P., Zhang, M., 2021. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. Nature Communications 12, 4828.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning (ICML), pp. 1–11.

Chen, X., Wang, X., Zhang, K., Fung, K.M., Thai, T.C., Moore, K., Mannel, R.S., Liu, H., Zheng, B., Qiu, Y., 2022a. Recent advances and clinical applications of deep learning in medical image analysis. Medical Image Analysis 79, 4.

Chen, Z., Li, G., Wan, X., 2022b. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge, in: Proceedings of the ACM International Conference on Multimedia, Association for Computing Machinery (ACM). pp. 5152–5161.

Decencière, E., Cazuguel, G., Zhang, X., Thibault, G., Klein, J.C., Meyer, F., Marcotegui, B., Quellec, G., Lamard, M., Danno, R., Elie, D., Massin, P., Viktor, Z., Erginay, A., Laÿ, B., Chabouis, A., 2013. Teleophta: Machine learning and image processing methods for teleophthalmology. IRBM 34, 196–203.

Decencière, E., Zhang, X., Cazuguel, G., Laÿ, B., Cochener, B., Trone, C., Gain, P., Ordóñez-Varela, J.R., Massin, P., Erginay, A., Charton, B., Klein, J.C., 2014. Feedback on a publicly distributed image database: The messidor database. Image Analysis and Stereology 33, 231–234.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8.

Derwin, D.J., Selvi, S.T., Singh, O.J., Shan, B.P., 2020. A novel automated system of discriminating microaneurysms in fundus images. Biomedical Signal Processing and Control 58, 101839.

Diaz-Pinto, A., Morales, S., Naranjo, V., Köhler, T., Mossi, J.M., Navea, A., 2019. Cnns for automatic glaucoma assessment using fundus images: An extensive validation. BioMedical Engineering Online 18.

Erhan, D., Manzagol, P.A., Bengio, Y., Bengio, S., Vincent, P., 2009. The difficulty of training deep architectures and the effect of unsupervised pre-training, in: Proceedings of the International Conference on Artificial Intelligence and Statistics (PMLR), pp. 153–160.

Eslami, S., de Melo, G., Meinel, C., 2021. Does clip benefit visual question answering in the medical domain as much as it does in the general domain?, in: Arxiv Preprint, pp. 1–9.

Fang, L., Wang, C., Li, S., Rabbani, H., Chen, X., Liu, Z., 2019. Attention to lesion: Lesion-aware convolutional neural network for retinal optical coherence tomography image classification. IEEE Transactions on Medical Imaging 38, 1959–1970.

Farnell, D.J., Hatfield, F.N., Knox, P., Reakes, M., Spencer, S., Parry, D., Harding, S.P., 2008. Enhancement of blood vessels in digital fundus photographs via the application of multiscale line operators. Journal of the Franklin Institute 345, 748–765.

Fauw, J.D., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., Hughes, C.O., Raine, R., Hughes, J., Sim, D.A., Egan, C., Tufail, A., Montgomery, H., Hassabis, D., Rees, G., Back, T., Khaw, P.T., Suleyman, M., Cornebise, J., Keane, P.A., Ronneberger, O., 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine 24, 1342–1350.

Finlayson, S.G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I.S., Saria, S., 2021. The clinician and dataset shift in artificial intelligence. The New England Journal of Medicine 385, 283–286.

Galdran, A., Dolz, J., Chakor, H., Lombaert, H., Ayed, I.B., 2020. Cost-sensitive regularization for diabetic retinopathy grading from eye fundus images, in: Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 1–7.

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y., 2021. Clip-adapter: Better vision-language models with feature adapters, in: ArXiv preprint, pp. 1–11. URL: `http://arxiv.org/abs/2110.04544`.

Garner, A., Ashton, N., 1979. Pathogenesis of hypertensive retinopathy: a review'. Journal of the Royal Society of Medicine 72.

Gass, M.J.D.M., 1988. Idiopathic senile macular hole its early stages and pathogenesis. Arch Ophthalmol. 106, 629–639.

Giancardo, L., Meriaudeau, F., Karnowski, T.P., Li, Y., Garg, S., Tobin, K.W., Chaum, E., 2012. Exudate-based diabetic macular edema detection in fundus images using publicly available datasets. Medical Image Analysis 16, 216–226.

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P.C., Mark, R., ., Stanley, H.E., 2000. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. Circulation 101, 215–220.

Goyal, S., Kumar, A., Garg, S., Raghunathan, Z.K.A., 2023. Finetune like you pretrain: Improved finetuning of zero-shot vision models, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 19338–19347.

Hamel, C., 2006. Retinitis pigmentosa. Orphanet Journal of Rare Diseases 1.

Hassan, T., Akram, M.U., Masood, M.F., Yasin, U., 2019. Deep structure tensor graph search framework for automated extraction and characterization of retinal layers and fluid pathology in retinal sd-oct scans. Computers in Biology and Medicine 105, 112–124.

Hassan, T., Akram, M.U., Werghi, N., Nazir, M.N., 2021. 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, 108–120.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–12.

Hoover, A., 2000. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. IEEE Transactions on Medical Imaging 19, 203–210.

Hoover, A., Goldbaum, M., 2003. Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels. IEEE Transactions on Medical Imaging 22, 951–958.

Hu, S.X., Li, D., Stühmer, J., Kim, M., Hospedales, T.M., 2022. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9068–9077.

Huang, J.H., Yang, C.H.H., Liu, F., Tian, M., Liu, Y.C., Wu, T.W., Lin, I.H., Wang, K., Morikawa, H., Chang, H., Tegner, J., Worring, M., 2021a. Deepopht: medical report generation for retinal images via deep models and visual explanation, in: Proceedings of the Winter Conference on Applications of Computer Vision (WACV), pp. 2442–2452.

Huang, S.C., Pareek, A., Jensen, M., Lungren, M.P., Yeung, S., Chaudhari, A.S., 2023. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. npj Digital Medicine 6.

Huang, S.C., Shen, L., Lungren, M.P., Yeung, S., 2021b. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3942–3951.

Jain, S., Agrawal, A., Saporta, A., Truong, S.Q., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., Langlotz, C.P., Rajpurkar, P., 2021. Radgraph: Extracting clinical entities and relations from radiology reports. NeurIPS: Track on Datasets

and Benchmarks .

Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T., 2021. Scaling up visual and vision-language representation learning with noisy text supervision, in: International Conference on Machine Learning (ICML), pp. 1–13.

Jin, K., Huang, X., Zhou, J., Li, Y., Yan, Y., Sun, Y., Zhang, Q., Wang, Y., Ye, J., 2022. Fives: A fundus image dataset for artificial intelligence based vessel segmentation. Scientific Data 9, 475.

Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., ying Deng, C., Mark, R.G., Horng, S., 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data 6.

Kanavati, F., Tsuneki, M., 2021. Partial transfusion: on the expressive influence of trainable batch norm parameters for transfer learning, in: MIDL, pp. 338–353.

Kauppi, T., Kalesnykiene, V., Kamarainen, J.K., Lensu, L., Sorri, I., Raninen, A., Voutilainen, R., Uusitalo, H., Kalviainen, H., Pietila, J., 2007. The diaretdb1 diabetic retinopathy database and evaluation protocol, in: Proceedings of the British Machine Vision Conference (BMVC), pp. 1–18.

Kovalyk, O., Morales-Sánchez, J., Verdú-Monedero, R., Sellés-Navarro, I., Palazón-Cabanes, A., Sancho-Gómez, J.L., 2022. Papila: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. Scientific Data 9, 291.

Krause, J., Gulshan, V., Rahimy, E., Karth, P., Widner, K., Corrado, G.S., Peng, L., Webster, D.R., 2018. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology 125, 1264–1272.

Kumar, A., Raghunathan, A., Jones, R.M., Ma, T., Liang, P., 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution, in: International Conference on Learning Representations (ICLR), pp. 1–42.

Kumar, J.R., Seelamantula, C.S., Gagan, J.H., Kamath, Y.S., Kuzhuppilly, N.I., Vivekanand, U., Gupta, P., Patil, S., 2023. Chaksu: A glaucoma specific fundus image database. Scientific Data 10.

Li, F., Song, D., Chen, H., Xiong, J., Li, X., Zhong, H., Tang, G., Fan, S., Lam, D.S., Pan, W., Zheng, Y., Li, Y., Qu, G., He, J., Wang, Z., Jin, L., Zhou, R., Song, Y., Sun, Y., Cheng, W., Yang, C., Fan, Y., Li, Y., Zhang, H., Yuan, Y., Xu, Y., Xiong, Y., Jin, L., Lv, A., Niu, L., Liu, Y., Li, S., Zhang, J., Zangwill, L.M., Frangi, A.F., Aung, T., yu Cheng, C., Qiao, Y., Zhang, X., Ting, D.S., 2020. Development and clinical deployment of a smartphone-based visual field deep learning system for glaucoma detection. npj Digital Medicine 3.

Li, L., Xu, M., Wang, X., Jiang, L., Liu, H., 2019a. Attention based glaucoma detection: A large-scale database and cnn model, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–10.

Li, T., Bo, W., Hu, C., Kang, H., Liu, H., Wang, K., Fu, H., 2021. Applications of deep learning in fundus images: A review. Medical Image Analysis1 69, 101971.

Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., Kang, H., 2019b. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Information Sciences 501, 511 – 522.

Lin, L., Li, M., Huang, Y., Cheng, P., Xia, H., Wang, K., Yuan, J., Tang, X., 2020. The sustech-sysu dataset for automated exudate detection and diabetic retinopathy grading. Scientific Data 7.

Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., Landman, B.A., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z., 2023. Clip-driven universal model for organ segmentation and tumor detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1–23.

Liu, R., Wang, X., Wu, Q., Dai, L., Fang, X., Yan, T., Son, J., Tang, S., Li, J., Gao, Z., Galdran, A., Poorneshwaran, J.M., Liu, H., Wang, J., Chen, Y., Porwal, P., Tan, G.S.W., Yang, X., Dai, C., Song, H., Chen, M., Li, H., Jia, W., Shen, D., Sheng, B., Zhang, P., 2022. Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. Patterns 3.

Lu, M.Y., Chen, B., Zhang, A., Williamson, D.F.K., Chen, R.J., Ding, T., Le, L.P., Chuang, Y.S., Mahmood, F., 2023. Visual language pretrained multiple instance zero-shot transfer for histopathology images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 19764–19775.

Matsoukas, C., Haslum, J.F., Sorkhei, M., Söderberg, M., Smith, K., 2022. What makes transfer learning work for medical images: Feature reuse and other factors, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9225–9234.

Menon, S., Vondrick, C., 2023. Visual classification via description from large language models, in: International Conference of Learning Representations (ICLR), pp. 1–17.

Moor, M., Banerjee, O., Abad, Z.S.H., Krumholz, H.M., Leskovec, J., Topol, E.J., Rajpurkar, P., 2023. Foundation models for generalist medical artificial intelligence. Nature 616, 259–265.

Müller, P., Kaissis, G., Zou, C., Rueckert, D., 2022. Joint learning of localized representations from medical images and reports, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 1–17.

Nakayama, L.F., Goncalves, M., Zago Ribeiro, L., Santos, H., Ferraz, D., Malerbi, F., Celi, L.A., Regatieri, C., 2023. A brazilian multilabel ophthalmological dataset (brset), in: PhysioNet, p. 1.

Neyshabur, B., Sedghi, H., Zhang, C., 2020. What is being transferred in transfer learning?, in: Advances in Neural Information Processing Systems (NeurIPS), pp. 1–12.

Niemeijer, M., Ginneken, B.V., Cree, M.J., Mizutani, A., Quellec, G., Sanchez, C.I., Zhang, B., Hornero, R., Lamard, M., Muramatsu, C., Wu, X., Cazuguel, G., You, J., Mayo, A., Li, Q., Hatanaka, Y., Cochener, B., Roux, C., Karray, F., Garcia, M., Fujita, H., Abramoff, M.D., 2010. Retinopathy online challenge: Automatic detection of microaneurysms in digital color fundus photographs. IEEE Transactions on Medical Imaging 29, 185–195.

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., Lee, J., Li, X., Liu, P., Lu, S., Murugesan, B., Naranjo, V., Phaye, S.S.R., Shankaranarayana, S.M., Sikka, A., Son, J., van den Hengel, A., Wang, S., Wu, J., Wu, Z., Xu, G., Xu, Y., Yin, P., Li, F., Zhang, X., Xu, Y., Zhang, X., Bogunović, H., 2019. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. Medical Image Analysis 59, 1–21.

Pachade, S., Porwal, P., Thulkar, D., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Giancardo, L., Quellec, G., Mériaudeau, F., 2021. Retinal fundus multi-disease image dataset (rfmid): A dataset for multi-disease detection research. Data 6, 1–14.

Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M., 2018. Radiology objects in context (roco): A multimodal image dataset, in: MICCAI Workshop: Large-scale Annotation of Biomedical Data and Expert Label Synthesis (LABELS), p. 180–189.

Pires, R., Jelinek, H.F., Wainer, J., Valle, E., Rocha, A., 2014. Advancing bag-of-visual-words representations for lesion classification in retinal images. PLoS ONE 9.

Porwal, P., Pachade, S., Kokare, M., Deshmukh, G., Son, J., Bae, W., Liu, L., Wang, J., Liu, X., Gao, L., Wu, T.B., Xiao, J., Wang, F., Yin, B., Wang, Y., Danala, G., He, L., Choi, Y.H., Lee, Y.C., Jung, S.H., Li, Z., Sui, X., Wu, J., Li, X., Zhou, T., Toth, J., Baran, A., Kori, A., Chennamsetty, S.S., Safwan, M., Alex, V., Lyu, X., Cheng, L., Chu, Q., Li, P., Ji, X., Zhang, S., Shen, Y., Dai, L., Saha, O., Sathish, R., Melo, T., Araújo, T., Harangi, B., Sheng, B., Fang, R., Sheet, D., Hajdu, A., Zheng, Y., Mendonça, A.M., Zhang, S., Campilho, A., Zheng, B., Shen, D., Giancardo, L., Quellec, G., Mériaudeau, F., 2020. Idrid: Diabetic retinopathy – segmentation

and grading challenge. Medical Image Analysis 59, 101561.

Qin, Z., Yi, H., Lao, Q., Li, K., 2023. Medical image understanding with pretrained vision language models: a comprehensive study, in: International Conference on Learing Representations (ICLR), pp. 1–20.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning (ICML), pp. 1–16.

Raghu, M., Zhang, C., Kleinberg, J., Bengio, S., 2019. Transfusion: Understanding transfer learning for medical imaging, in: Advances in neural information processing systems (NeurIPS), pp. 1–11.

Ruiz-Medrano, J., Montero, J.A., Flores-Moreno, I., Arias, L., García-Layana, A., Ruiz-Moreno, J.M., 2019. Myopic maculopathy: Current status and proposal for a new classification and grading system (atn). Progress in Retinal and Eye Research 69, 80–115.

Sengupta, S., Singh, A., Leopold, H.A., Gulati, T., Lakshminarayanan, V., 2020. Ophthalmic diagnosis using deep learning with fundus images – a critical review. Artificial Intelligence in Medicine 102, 101758.

Shu, M., Nie, W., Huang, D.A., Yu, Z., Goldstein, T., Anandkumar, A., Xiao, C., 2022. Test-time prompt tuning for zero-shot generalization in vision-language models. Advances in Neural Information Processing Systems 35, 14274–14289.

Sivaswamy, J., Krishnadas, S.R., Joshi, G.D., Jain, M., Tabish, A.U.S., 2014. Drishti-gs retinal image dataset for optic nerve head segmentation, in: International Symposium on Biomedical Imaging (ISBI), pp. 53–56.

Sun, R., Li, Y., Zhang, T., Mao, Z., Wu, F., Zhang, Y., 2021. Lesion-aware transformers for diabetic retinopathy grading, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10938–10939.

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2017. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE Transactions on Medical Imaging 35, 1299–1312.

Takahashi, H., Tampo, H., Arai, Y., Inoue, Y., Kawashima, H., 2017. Applying artificial intelligence to disease staging: Deep learning for improved staging of diabetic retinopathy. PLoS ONE 12.

Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P., 2022. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature Biomedical Engineering .

de Vente, C., Vermeer, K.A., Jaccard, N., Wang, H., Sun, H., Khader, F., Truhn, D., Aimyshev, T., Zhanibekuly, Y., Le, T.D., Galdran, A., Gonzalez Ballester, M.A., Carneiro, G., G, D.R., S, H.P., Puthussery, D., Liu, H., Yang, Z., Kondo, S., Kasai, S., Wang, E., Durvasula, A., Heras, J., Zapata, M.A., Araujo, T., Aresta, G., Bogunovic, H., Arikan, M., Lee, Y.C., Cho, H.B., Choi, Y.H., Qayyum, A., Razzak, I., van Ginneken, B., Lemij, H.G., Sanchez, C.I., 2023. Airogs: Artificial intelligence for robust glaucoma screening challenge. ArXiv preprint .

Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., Yu, L., 2022a. Multi-granularity cross-modal alignment for generalized medical visual representation learning, in: Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), pp. 1–14.

Wang, Z., Wu, Z., Agarwal, D., Sun, J., 2022b. Medclip: Contrastive learning from unpaired medical images and text, in: Empirical Methods in Natural Language Processing (EMNLP), pp. 1–12.

Wei, Q., Li, X., Wang, H., Ding, D., Yu, W., Chen, Y., 2018. Laser scar detection in fundus images using convolutional neural networks, in: Asian Conference on Computer Vision (ACCV), pp. 191–206.

WHO, 2019. World report of vision. World Health Organization .

Wilkinson, C.P., Ferris, F.L., Klein, R.E., Lee, P.P., Agardh, C.D., Davis, M., Dills, D., Kampik, A., Pararajasegaram, R., Verdaguer, J.T., Lum, F., 2003. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. Ophthalmology 110, 1677–1682.

Windsor, R., Jamaludin, A., Kadir, T., Zisserman, A., 2023. Vision-language modelling for radiological imaging and reports in the low data regime, in: Medical Image with Deep Learning (MIDL), pp. 1–21.

Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Gontijo-Lopes, R., Hajishirzi, H., Farhadi, A., Namkoong, H., Schmidt, L., 2022. Robust fine-tuning of zero-shot models, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7959–7971.

Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W., 2023. Medklip: Medical knowledge enhanced language-image pre-training in radiology, in: ArXiv Preprint, pp. 1–16. URL: `http://arxiv.org/abs/2301.02228`.

Wójcik, M.A., 2022. Foundation models in healthcare: Opportunities, biases and regulatory prospects in europe. EGOVIS 13429 LNCS, 32–46.

Xiaomeng, L., Hu, X., Lequan, Y., Zhu, L., Fu, C.W., Heng, P.A., 2020. Canet: Cross-disease attention network for joint diabetic retinopathy and diabetic macular edema grading. IEEE Transactions on Medical Imaging 5, 1483–1494.

Xie, X., Niu, J., Member, S., Liu, X., Chen, Z., Tang, S., Yu, S., 2021. A survey on incorporating domain knowledge into deep learning for medical image analysis. Medical Image Analysis 69, 101985.

Yang, J., Li, C., Zhang, P., Xiao, B., Liu, C., Yuan, L., Gao, J., 2022. Unified contrastive learning in image-text-label space, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), pp. 19163–19173.

Yang, J.J., Li, J., Shen, R., Zeng, Y., He, J., Bi, J., Li, Y., Zhang, Q., Peng, L., Wang, Q., 2016. Exploiting ensemble learning for automatic cataract detection and grading. Computer Methods and Programs in Biomedicine 124, 45–57.

Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H., 2022a. Tip-adapter: Training-free clip-adapter for better vision-language modeling, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 1–19.

Zhang, S., Xu, Y., Usuyama, N., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Lungren, M., Naumann, T., Poon, H., 2023. Large-scale domain-specific pretraining for biomedical vision-language processing. URL: `https://arxiv.org/abs/2303.00915`, doi:`10.48550/ARXIV.2303.00915`.

Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P., 2022b. Contrastive learning of medical visual representations from paired images and text, in: Machine Learning for Healthcare (MLHC), pp. 1–16.

Zhang, Z., Yin, F.S., Liu, J., Wong, W.K., Tan, N.M., Lee, B.H., Cheng, J., Wong, T.Y., 2010. Origa-light: An online retinal fundus image database for glaucoma analysis and research, in: Annual International Conference of the IEEE Engineering in Medicine and Biology, pp. 3065–3068.

Zhao, S., Zhang, Z., Schulter, S., Zhao, L., Vijay Kumar, B., Stathopoulos, A., Chandraker, M., Metaxas, D.N., 2022. Exploiting unlabeled data with vision and language models for object detection, in: European Conference on Computer Vision, pp. 159–175.

Zhao, Z., Zhang, K., Hao, X., Tian, J., Chua, M.C.H., Chen, L., Xu, X., 2019. Bira-net bilinear attention net for diabetic retinopathy grading, in: International Conference on Image Processing (ICIP), pp. 1385–1389.

**Supplementary Materials.**
**A Foundation LAnguage-Image model of the Retina (FLAIR): Encoding expert knowledge in text supervision.**

## S1. Dataset details

This section provides supplementary details regarding the assembly of datasets for the foundation fundus model training. Also, it includes supplementary datasets and partitions used during the evaluation stage, as well as visual descriptions of the categories selected to evaluate the proposed foundation model on novel categories.

***Categories***. In the following, we provide the categories and corresponding abbreviations used for training and testing the proposed foundation model. No diabetic retinopathy (noDR), mild diabetic retinopathy (mildDR), moderate diabetic retinopathy (modDR), severe diabetic retinopathy (sevDR), proliferative diabetic retinopathy (prolDR), noisy, clean, diabetic macular edema (DME), no referable diabetic macular edema (noDME), hard exudate (hEX), soft exudate (sEX), microaneurysms (MA), haemorrhages (HE), non clinically significant diabetic macular edema (nonCSDME), age-related macular degeneration (ARMD), media haze (MH), drusen (DN), pathologic myopia (MYA), branch retinal vein occlusion (BRVO), tessellation (TSLN), epiretinal membrane (ERM), laser scar (LS), macular scar (MS), central serous retinopathy (CSR), optic disc cupping (ODC), central retinal vein occlusion (CRVO), tortuous vessels (TV), asteroid hyalosis (AH), optic disc pallor (ODP), optic disc edema (ODE), shunt (ST), anterior ischemic optic neuropathy (AION), parafoveal telangiectasia (PT), retinal traction (RT), retinitis (RS), chorioretinitis (CRS), exudate (EX), retinal pigment epithelium changes (RPEC), macular hole (MHL), retinitis pigmentosa (RP), cotton wool spots (CWS), colobomas (CB), optic disc pit maculopathy (ODM), preretinal haemorrhage (PRH), myelinated nerve fibers (MNF), haemorrhagic retinopathy (HR), central retinal artery occlusion (CRAO), tilted disc (TD), cystoid macular edema (CME), post traumatic choroidal rupture (PTCR), choroidal folds (CF), vitreous haemorrhage (VH), macroaneurysm (MCA), vasculitis (VS), branch retinal artery occlusion (BRAO), plaque (PLQ), haemorrhagic pigment epithelial detachment (HPED), collateral (CL), normal (N), large optic cup (LOC), retina detachment (RD), Vogt-Koyanagi syndrome (VKH), maculopathy (M), Glaucoma (G), optic atrophy (OA), severe hypertensive retinopathy (sevHR), disc swelling and elevation (DSE), dragged disk (DD), congenital disk abnormality (CDA), Bietti crystalline dystrophy (BCD), peripheral retinal degeneration and break (PRDB), neoplasm (NP), yellow-white spots flecks (YWSF), fibrosis (F), silicon oil (SO), no proliferative diabetic retinopathy (noProlDR), no glaucoma (noG), cataract (CAT), hypertensive retinopathy (HR), neovascular age-related macular degeneration (neovARMD), geographical age-related macular degeneration (geoARMD), acute central serous retinopathy (acCSR), chronic central serous retinopathy (chCSR), no cataract (noCAT), abnormal optic disc (AOD), abnormal vessels (AV), abnormal macula (AM), macular edema (ME), scar (S), nevus (NE), increased cup disk (ICD), intraretinal microvascular abnormalities (IrMA), red small dots (ReSD), neovascularization (neoV), disease (Dis), superficial haemorraghe (supHE), deep haemorraghe (deepHE).

***Labels distribution***. We depict in Figure S1 the label distribution of the assembled dataset, used for training the universal fundus model.

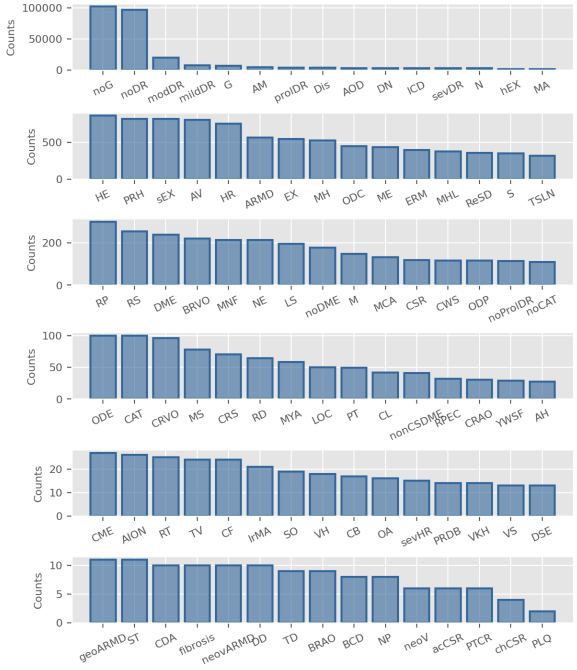



Figure S1: **Distribution of the categories in the dataset assembly**. The plot depicts the number of samples within each category in the assembly dataset, ordered from the most represented class to the least. One may observe a long-tail distribution, which is common when integrating multiple public medical-imaging datasets (Liu et al., 2023).

***Unseen categories***. Figure S2 visualizes examples of the target categories selected to explore the transferability of FLAIR to unseen diseases. Also, we present domain-knowledge descriptors generated for prompt-based classification.

***Supplementary validation partitions***. As stated in the main manuscript, the categories cataract, pathological myopia, retinitis pigmentosa, and macular hole are used for evaluating the transferability of the universal model to downstream tasks. Thus, we proposed two different partitions by retrieving samples from specific datasets of the database assembly. First, ODIR200x3 contains 200 samples for normal fundus images and 200 diagnosed with cataract and pathological myopia, which are sampled from the ODIR-5K dataset. Second,

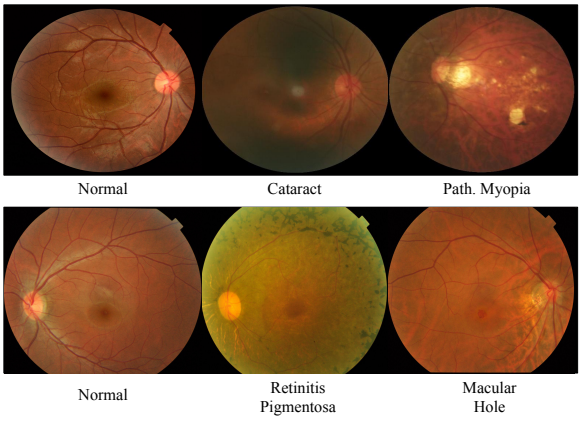



Figure S2: **Novel categories for transferability evaluation.** Visualization of the different diseases used for validating the capabilities of the foundation model for adaptation to novel categories. Samples obtained from the ODIR-5K (*top row*) and 1000x39 (*bottom row*) datasets. Using expert's domain knowledge on retinal fundus image analysis, we define the following findings as descriptors of the different diseases: cataracts are featured ''*opacities in the macular area*" (Allen and Vasavada, 206; Yang et al., 2016), pathological myopia as ''*anomalous disc, macular atrophy, and possible tessellation*" (Ruiz-Medrano et al., 2019), retinitis pigmentosa is characterized for ''*pigment deposits are present in the periphery*" (Hamel, 2006), and a macular hole might be described as ''*grayish fovea*"/''*lesion in the macula*" (Gass, 1988).

20x3 contains 20 normal examples and 20 examples of retinitis pigmentosa and macular hole, which are sampled from the 1000x39 dataset. The target diseases were not used during the training of the foundation models. Thus, we create supplementary evaluation datasets by using the discarded samples of the other datasets. First, we create CAT-MYA-2 partition by retrieving samples with cataract from the Cataract dataset and samples with pathological myopia from the BRSET dataset. Since this subset is designed to validate adapters trained on the ODIR200x3 partition, we select normal samples from the other evaluation subset, 20x3. We followed the same procedure to create the evaluation partition RP-MHL-2, consisting of normal samples from ODIR200x3 and retinitis pigmentosa and macular hole samples from RFMid dataset. We present a summary of supplementary partitions in Table S1. It is worth mentioning that the ensuing partitions present a small number of samples, and are used only to evaluate the generalization capability of adapters trained on new tasks, under the low-data regime. In addition, we also include the ACRIMA and DeepDRiD datasets for the additional validation of the proposed methods for generalization and adaptation to glaucoma detection and DR grading, respectively .

## S2. Results

This section provides supplementary results to support the capabilities of the proposed universal model. In particular, we first present results on zero-shot classification for novel categories on the supplementary evaluation datasets. Secondly, we introduce the results obtained for the transferability of the pre-trained foundation model to additional datasets.

Table S1: **Supplementary datasets for the evaluation of the foundation model**. We set aside additional datasets for Glaucoma and DR grading evaluation, ACRIMA and DeepDRiD, respectively. Also, we combine samples corresponding to the categories selected as unseen for FLAIR training (i.e., RP, MHL, CAT, and MYA), which are discarded during training.

| Dataset | #Images | Labels |
|---|---|---|
| *Domain shift* | | |
| DeepDRiD | 2,256 | noDR, mildDR, modDR, sevDR, prolDR. |
| ACRIMA | 705 | G, noG |
| *Unseen categories* | | |
| RP-MHL-2 | 30 | N, RP, MHL |
| CAT-MYA-2 | 60 | N, CAT, MYA |

### *S2.1. Performance on novel classes*

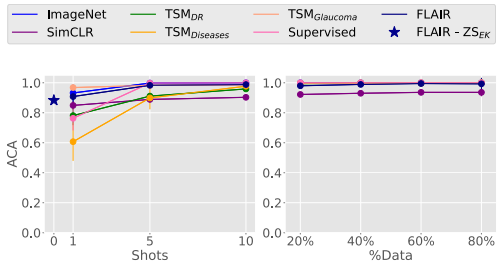
***Zero-shot classification***. We present in Table S2 the results obtained regarding prompt-based classification on unseen categories (i.e. no adaptation) for the supplementary datasets RP-MHL-2 and CAT-MYA-2. The obtained results are in line with the ones observed in the main paper. Using domain-knowledge descriptors for prompt augmentation during training outperforms the base CLIP trained on medical images, and using well-designed prompts instead of target-category names for inference provides remarkable improvements, of $\sim$ 4% and $\sim$ 11%, respectively. Also, in the CAT-MYA-2 dataset, the base CLIP model trained on natural images is able to obtain promising results - in contrast to the other three partitions used for zero-shot evaluation. It should be noted that CLIP is trained on 400M of images and text pairs, including some medical imaging datasets (Radford et al., 2021).

Table S2: **Zero-shot classification on supplementary datasets.** Transferability of the proposed foundation model to new tasks via prompt-based inference. Three different strategies to generate text prompts are evaluated: anomaly detection (i.e., ''*normal/disease*"), classification via naive prompt (i.e., the new disease name) or designed prompts (i.e., domain-knowledge descriptors). The metric presented is the accuracy for each category. The proposed method, FLAIR-$\pi_{EK}$, is shadowed, whereas the best results are highlighted in bold.

| Method | Dataset | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | RP-MHL-2 | | | | CAT-MYA-2 | | | |
| | N | RP | MHL | Avg. | N | CAT | MYA | Avg. |
| *Anomaly Detection Inference (i.e. "normal/disease")* | | | | | | | | |
| CLIP | 0.800 | 0.050 | | 0.425 | 1.000 | 0.075 | | 0.538 |
| BiomedCLIP | 0.900 | 0.150 | | 0.525 | 0.950 | 0.375 | | 0.662 |
| FLAIR-$\pi_{naive}$ | 1.000 | 1.000 | | **1.000** | 0.800 | 0.050 | | 0.425 |
| FLAIR-$\pi_{EK}$ | 0.900 | 1.000 | | 0.950 | 0.850 | 0.500 | | **0.675** |
| *Inference with Naive Prompts - $\pi_{naive}$ (e.g. "cataract")* | | | | | | | | |
| CLIP | 0.000 | 0.600 | 0.100 | 0.233 | 1.000 | 0.750 | 0.050 | **0.600** |
| BiomedCLIP | 0.800 | 0.900 | 0.100 | 0.600 | 0.950 | 0.500 | 0.100 | 0.517 |
| FLAIR-$\pi_{naive}$ | 1.000 | 0.000 | 0.700 | 0.567 | 0.450 | 0.750 | 0.000 | 0.400 |
| FLAIR-$\pi_{EK}$ | 0.900 | 0.500 | 0.600 | **0.667** | 0.950 | 0.550 | 0.000 | 0.500 |
| *Inference with Expert Knowledge Prompts - $\pi_{EK}$ (e.g. "opacity in the macular area")* | | | | | | | | |
| CLIP | 1.000 | 0.000 | 0.000 | 0.333 | 0.900 | 0.250 | 0.950 | **0.700** |
| BiomedCLIP | 0.500 | 0.100 | 0.600 | 0.400 | 0.200 | 0.450 | 1.000 | 0.550 |
| FLAIR-$\pi_{naive}$ | 1.000 | 0.400 | 0.000 | 0.467 | 0.850 | 0.800 | 0.050 | 0.567 |
| FLAIR-$\pi_{EK}$ | 1.000 | 1.000 | 0.300 | **0.767** | 1.000 | 0.750 | 0.000 | 0.583 |

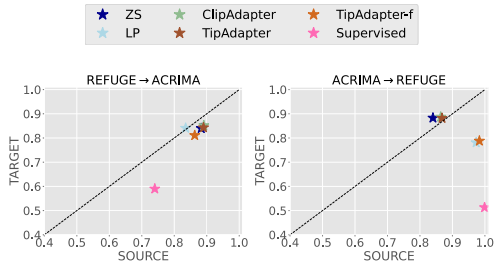
## S2.2. *Transferability*

***Transferability for Glaucoma detection***. Figure S3 shows the transferability results for glaucoma detection on the ACRIMA dataset. In addition, we present in Figure S3 the generalization performance of adapters, using the REFUGE dataset. Concretely, the adapters are adjusted on one of the datasets, and evaluated on both of them. This process is carried out crosswise, under the low-data regime.




ACRIMA - *Domain shift - Glaucoma*

Figure S3: **Transferability on ACRIMA dataset.** Results of transferring the feature representations of the pre-trained models to ACRIMA dataset for glaucoma detection in the low-data (*left column*) and large-data (*right column*) regimes. The results were obtained by adjusting a linear-probe classifier. The metric presented is the average accuracy, averaged across 5 cross-validation folds. ZS: zero-shot (i.e. prompt-based classification).

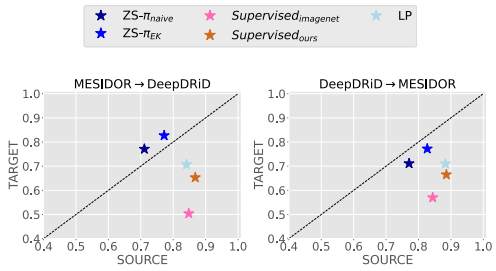



*Domain shift - Glaucoma*

Figure S4: **Generalization of the adapters for glaucoma detection.** Evaluation of the generalization capabilities of the few-shot adapters under domain shifts. The adapters are tuned on the source domain, and the performance is evaluated on both the source and target datasets. The metric presented is the average accuracy, averaged across 5 cross-validation folds. The results were obtained using $k = 10$ shots.

Regarding transferability, the obtained results are in line with those observed in the other datasets, which we provided in the main paper. First, the foundation model, pre-trained using language supervision on a large number of tasks, shows better transferability than task-specific models, including the one trained for the glaucoma detection task. Secondly, it performs on par with the supervised, dataset-specific model, while requiring less numbers of shots, with only the linear-probe classifier being tuned, in an efficient way. Still, in this dataset, in contrast to the other evaluation partitions, the supervised counterpart reaches almost a perfect performance while requiring very few shots. Nevertheless, the resultant model shows poor transferability under domain shift when tested on the REFUGE dataset (see Figure S4). In contrast, vision-language adapters over the foundation model tuned on the source domain are able to maintain the performance on the target set. Interestingly, zero-shot classification (i.e., no adaptation) shows the best transferrability between both subsets. It is worth mentioning that ACRIMA combines different centers to obtain glaucoma and non-glaucoma fundus samples, which might introduce a bias when training dataset-specific models.

***Transferability for DR grading***. In the following, we present experiments to study the generalization capabilities of the pre-trained foundation model for DR grading. Concretely, we use the DeepDRiD challenge (Liu et al., 2022) dataset for this task. In this section, we evaluate the model performance via the quadratic Cohen kappa, the main figure of merit used for DR grading evaluation. Figure S5 presents the generalization performance of different strategies tuned under the large-data regime on DeepDRiD and MESIDOR, evaluated on the source domain, and on the other dataset, used as a target. The evaluated methods are prompt-based classification using the foundation model via category names ($\pi_{naive}$), using an ensemble of expert knowledge descriptions for each DR class ($\pi_{EK}$), and a linear-probe adapter. In addition, we train dataset-specific models (i.e., full fine-tuning) using ResNet-50 initialized on ImageNet (Supervised$_{Imagenet}$), and on the foundation fundus model (Supervised$_{ours}$).




*Domain shift - DR*

Figure S5: **Adaptation stage generalization for DR grading.** Evaluation of the generalization capabilities of adapters and dataset-specific models. Different strategies are tuned on the source domain, and the performance is evaluated on both source and target datasets. The metric presented is the quadratic Cohen kappa ($\kappa$). Results were obtained on the large data regime, using the whole train subset from DeepDiD, and 80% of MESIDOR samples for training, respectively.

The obtained results using full fine-tuning resemble the outstanding performance obtained in the DeepDRiD challenge, in which leaderboard methods reached $[0.90, 0.93]$ quadratic kappa. It is worth mentioning that due to the heuristics for

ranking optimization common to these competitions, these methods include additional strategies such as test-time adaptation, model ensemble, or selection of the most appropriate backbone. Thus, they may reach slightly better performance on the source domain, but no cross-domain performance is assessed. Nonetheless, common conclusions can be drawn: using a model with initialized weights on pre-trained fundus analysis tasks yields improved results. However, our results suggest that these source-domain specialized models fail remarkably when faced with a domain shift. In this context, the proposed foundation model presents interesting properties. First, by simply fitting a linear-probe classifier, the model achieves a performance on par with the full-backbone fine-tuning, while not being penalized as much in the scenario of a domain shift. Second, prompt-based classification with no adaptation reaches the best out-of-distribution performance, which improves using an ensemble of domain-knowledge descriptors for both domains.

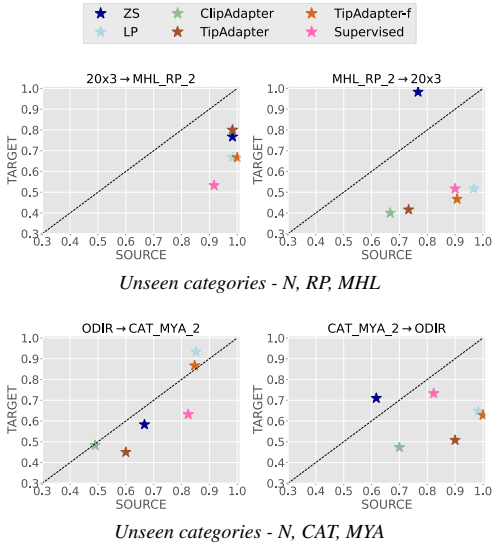



Figure S6: **Adapter generalization.** Evaluation of the generalization capabilities of the few-shot adapters for novel classes under domain shift. Adapters are tuned on the source domain, and the performance is evaluated on another dataset with the same categories. The metric presented is the average accuracy, averaged across 5 cross-validation folds. The results were obtained using $k = 5$ shots.

***Generalization of the few-shot adapters on unseen categories***. In the main experiments of this paper, we have empirically presented the limitations of the few-shot vision-language adapters in the medical context, with respect to domain knowledge prompts (if $k \leq 5$) or linear probes of vision features (if $k \geq 10$); see Section 5.2.2 for more details. Still, the linear-probe adapter seems to generalize worse against sequential domain shifts after the tuning stage, compared to zero-shot classification using text prompts (see Section 5.2.3). This motivates additional experiments to validate the effectiveness of vision-language adapters in this setting. In particular, we evaluate the performance of the adapters for unseen categories in the foundation model, tuned on a source domain, and evaluated on both, source and target domains. Experiments are carried out in the low-data regime, to reproduce a realistic scenario for vision-language adapters, where only a handful of labeled samples are available. Results are depicted in Figure S6. As already presented, the adapters struggle to reach the best performance in the source domain. Nevertheless, these adapters are able to maintain a more robust performance on the target domains (see Figure S6, *bottom-right* and *top-right*), or even slight improvements for both, source and target (see Figure S6, TipAdapter *top-left*, TipAdapter-f *bottom-left*).

Table S3: **Expert Knowledge descriptions.**

| Category | Domain Knowledge descriptor |
|---|---|
| no diabetic retinopathy | ”no relevant haemorrhages, microaneurysms or exudates” / ”no microaneurysms” / ”no referable lesions” |
| mild diabetic retinopathy | ”few microaneurysms” / ”few hard exudates” / ”few retinal haemorrhages” |
| moderate diabetic retinopathy | ”retinal haemorrhages in few quadrants” / ”many haemorrhages” / ”cotton wool spots” |
| severe diabetic retinopathy | ”severe haemorrhages in all four quadrants” / ”venous beading” / ”intraretinal microvascular abnormalities” |
| proliferative diabetic retinopathy | ”diabetic retinopathy with neovascularization at the disk” / ”neovascularization” |
| diabetic macular edema | ”macular edema” / ”presence of exudates” / ”leakage of fluid within the central macula from microaneurysms” / ”presence of exudates within the radius of one disc diameter from the macula center” |
| no referable diabetic macular edema | ”no apparent exudates” |
| hard exudates | ”small white or yellowish deposits with sharp margins” / ”bright lesion” |
| soft exudates | ”pale yellow or white areas with ill-defined edges” / ”cotton-wool spot” / ”small, whitish or grey, cloud-like, linear or serpentine, slightly elevated lesions with fimbriated edges” |
| microaneurysms | ”small red dots” |
| haemorrhages | ”dense, dark red, sharply outlined lesion” |
| non clinically significant diabetic macular edema | ”presence of exudates outside the radius of one disc diameter from the macula center” / ”presence of exudates” |
| age-related macular degeneration | ”many small drusen” / ”few medium-sized drusen” / ”large drusen” |
| media haze | ”vitreous haze” / ”pathological opacity” / ”the obscuration of fundus details by vitreous cells and protein exudation” |
| drusens | ”yellow deposits under the retina” / ”numerous uniform round yellow-white lesions” |
| pathologic myopia | ”tilted disc, peripapillary atrophy, and macular atrophy. There are chorioretinal scars in the inferonasal periphery” / ”maculopahy” |
| branch retinal vein occlusion | ”occlusion of one of the four major branch retinal veins” |
| tessellation | ”large choroidal vessels at the posterior fundus” |
| epiretinal membrane | ”greyish semi-translucent avascular membrane” |
| laser scar | ”round or oval, yellowish-white with variable black pigment centrally” / ”50 to 200 micron diameter lesions” |
| central serous retinopathy | ”subretinal fluid involving the fovea” / ”leakage” |
| asteroid hyalosis | ”multiple sparking, yellow-white, and refractile opacities in the vitreous cavity” / ”vitreous opacities” |
| optic disc pallor | ”pale yellow discoloration that can be segmental or generalized on optic disc” |
| shunt | ”collateral vessels connecting the choroidal and the retinal vasculature” / ”collateral vessels of large caliber and lack of leakage” |
| exudates | ”small white or yellowish-white deposits with sharp margins” / ”bright lesion” |
| macular hole | ”a lesion in the macula” / ”small gap that opens at the centre of the retina” |
| retinitis pigmentosa | ”bone spicule-shaped pigment deposits are present in the mid periphery” / ”retinal atrophy” ”the macula is preserved” / ”peripheral ring of depigmentation” / ”arteriolar attenuation and atrophy of the retinal pigmented epithelium” |
| cotton wool spots | ”soft exudates” |
| glaucoma | ”optic nerve abnormalities” / ”abnormal size of the optic cup” / ”anomalous size in the optic disc” |
| severe hypertensive retinopathy | ”flame-shaped hemorrhages at the disc margin, blurred disc margins” / ”congested retinal veins, papilledema, and secondary macular exudates” / ”arterio-venous crossing changes, macular star and cotton wool spots” |
| no proliferative diabetic retinopathy | ”diabetic retinopathy with no neovascularization” / ”no neovascularization” |
| hypertensive retinopathy | ”possible signs of hemorrhage with blot, dot, or flame-shaped” / ”possible presence of microaneurysm, cotton-wool spot, or hard exudate” / ”arteriolar narrowing” / ”vascular wall changes” / ”optic disk edema” |
| intraretinal microvascular abnormalities | ”shunt vessels and appear as abnormal branching or dilation of existing blood vessels (capillaries) within the retina” / ”deeper in the retina than neovascularization, has blurrier edges, is more of a burgundy than a red, does not appear on the optic disc” / ”vascular loops confined within the retina” |
| red small dots | ”microaneurysms” |
| a disease | ”no healthy” / ”lesions” |
| normal | ”healthy” / ”no findings” / ”no lesion signs” |