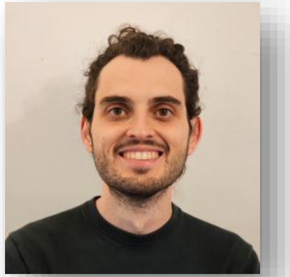
Towards Multi-Modal Foundation Models for Retinal Image Analysis

Julio Silva-Rodríguez



Jose Dolz

Ismail Ben Ayed

ÉTS Montréal

Hadi Chakor MD

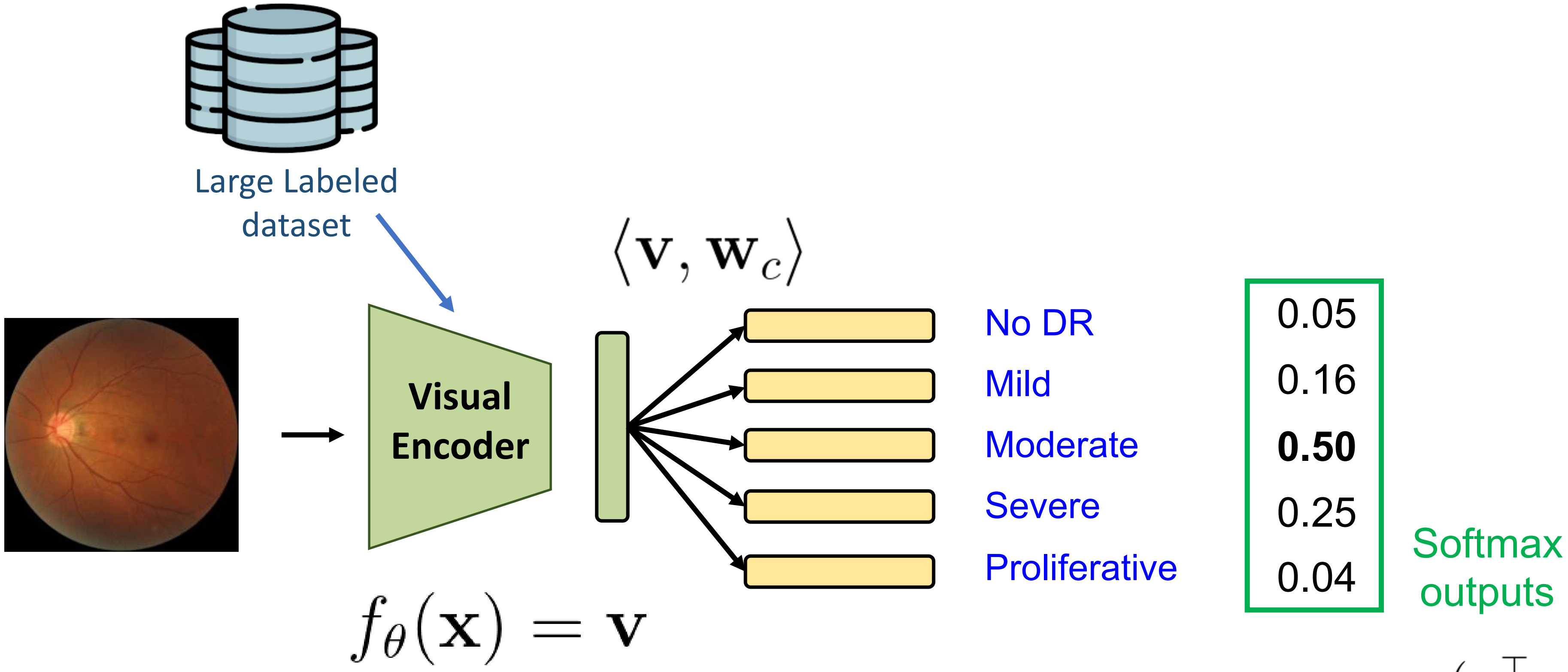
Riadh Kobbi

DIAGNOS Inc.

Financial Disclosure

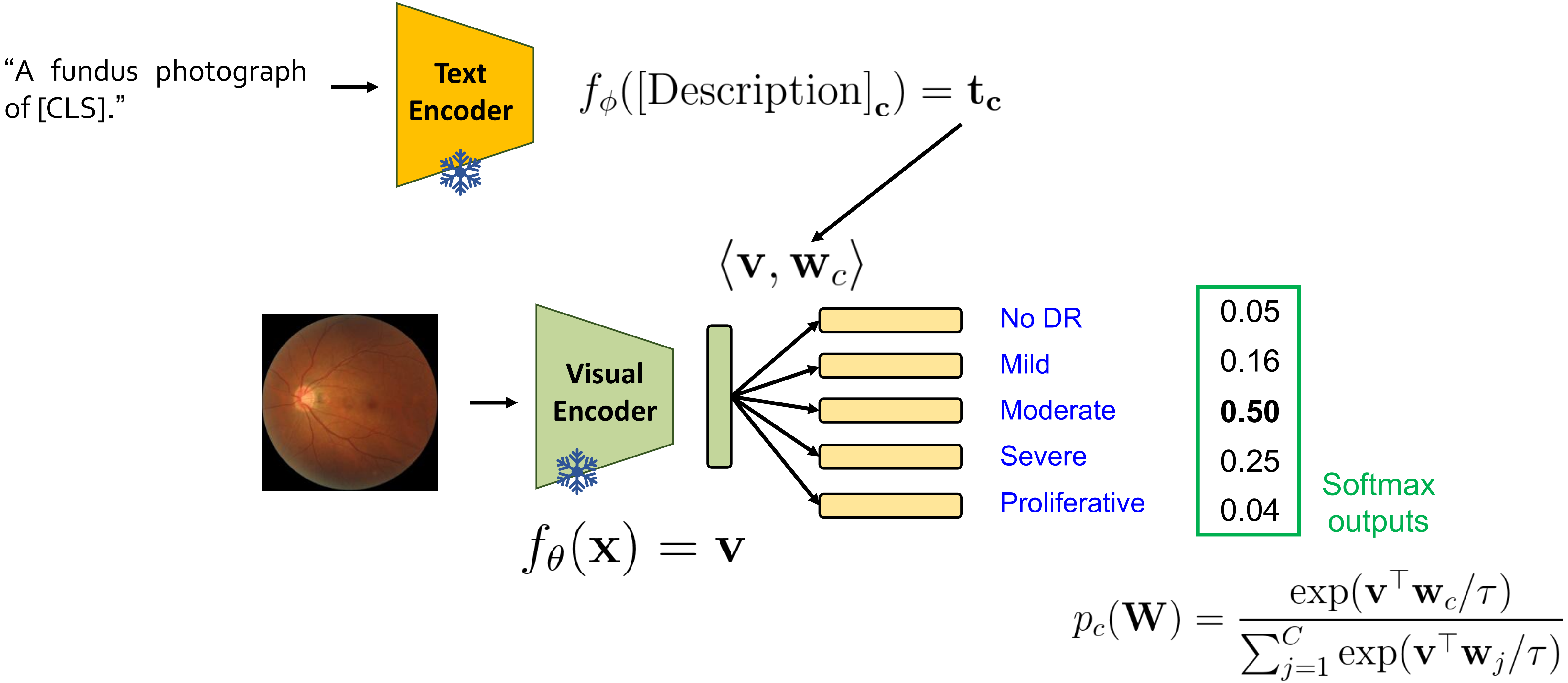
I **do not have** any affiliation (financial or otherwise) with a commercial organization that may have a direct or indirect connection to the content of my presentation.

Classical dataset-specific models

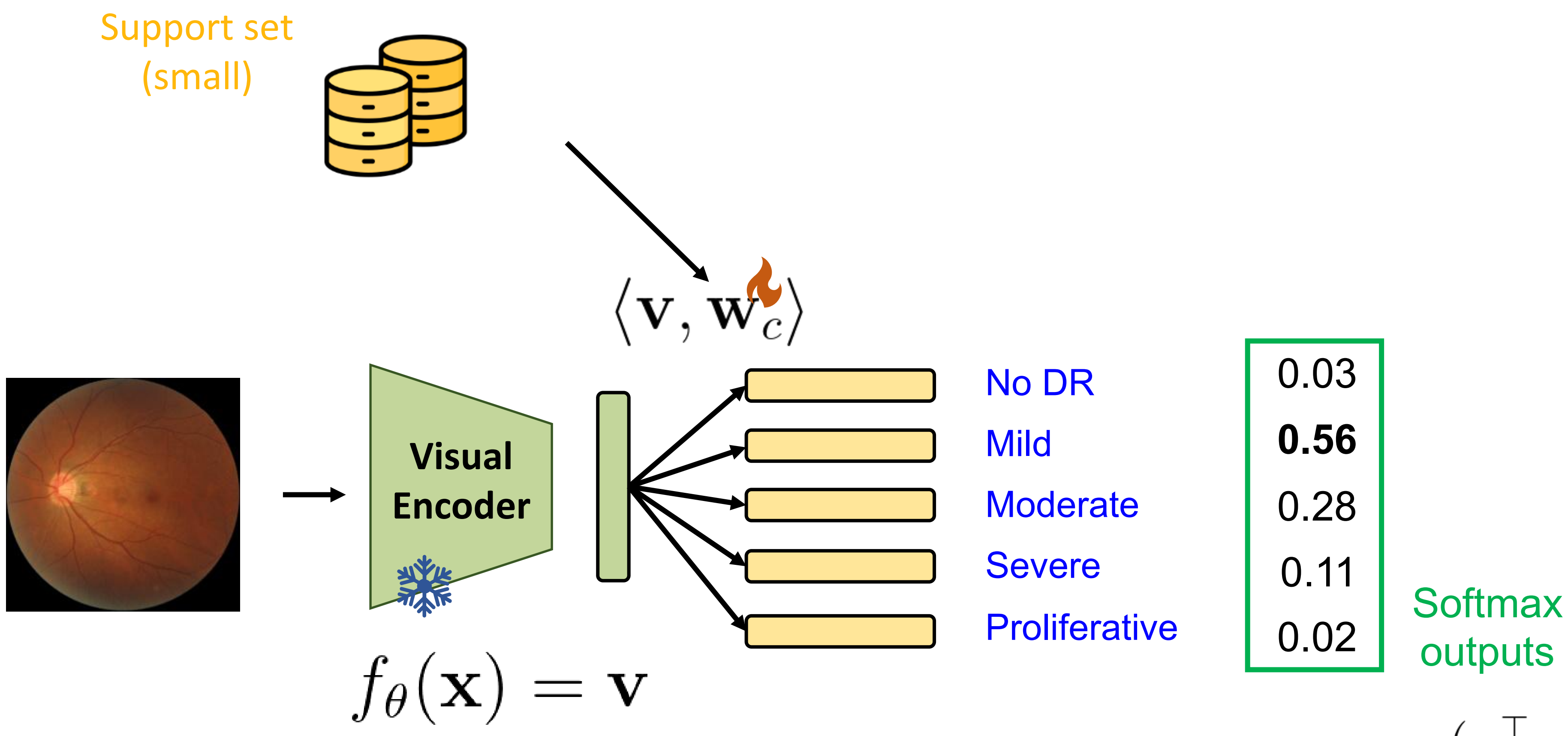


$$p_c(\mathbf{W}) = \frac{\exp(\mathbf{v}^\top \mathbf{w}_c / \tau)}{\sum_{j=1}^{C} \exp(\mathbf{v}^\top \mathbf{w}_j / \tau)}$$

Vision-Language Models

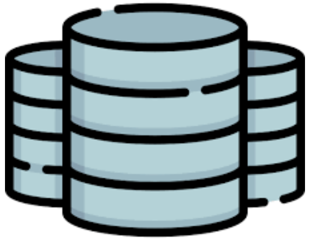


Efficient, linear probe transfer

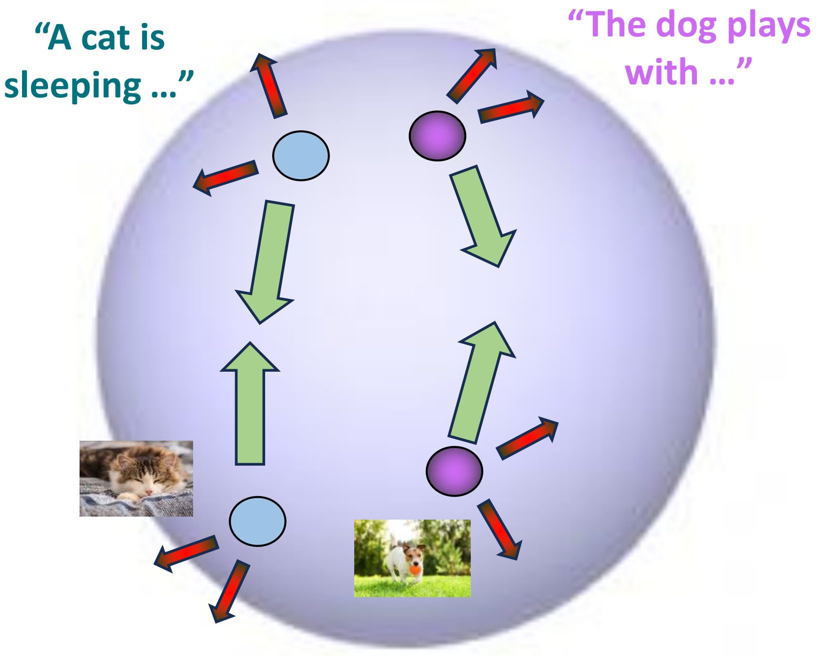


$$p_c(\mathbf{W}) = \frac{\exp(\mathbf{v}^\top \mathbf{w}_c / \tau)}{\sum_{j=1}^{C} \exp(\mathbf{v}^\top \mathbf{w}_j / \tau)}$$

How are VLMs pre-trained?



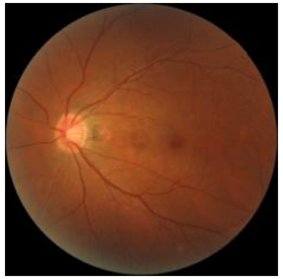
(very large) Image-text pairs, e.g., 400M CLIP



"A fundus photograph of [CLS]." → **Text Encoder**

$$f_\phi([\text{Description}]_{\mathbf{c}}) = \mathbf{t_c}$$

$$\langle \mathbf{v}, \mathbf{w}_c \rangle$$

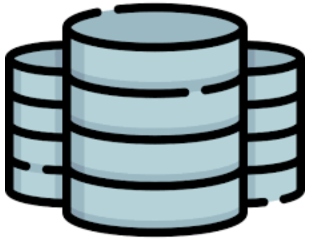


→ **Visual Encoder**

No DR	0.05
Mild	0.16
Moderate	**0.50**
Severe	0.25
Proliferative	0.04

$$f_\theta(\mathbf{x}) = \mathbf{v}$$

How are VLMs pre-trained?

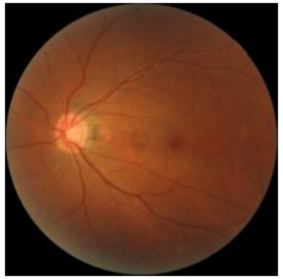


(very large)
Image-text pairs,
e.g., 400M CLIP

"A fundus photograph of [CLS]." → **Text Encoder**

$$f_\phi([\text{Description}]_{\mathbf{c}}) = \mathbf{t_c}$$

$$\langle \mathbf{v}, \mathbf{w}_c \rangle$$



→ **Visual Encoder**

No DR	0.05
Mild	0.16
Moderate	**0.50**
Severe	0.25
Proliferative	0.04

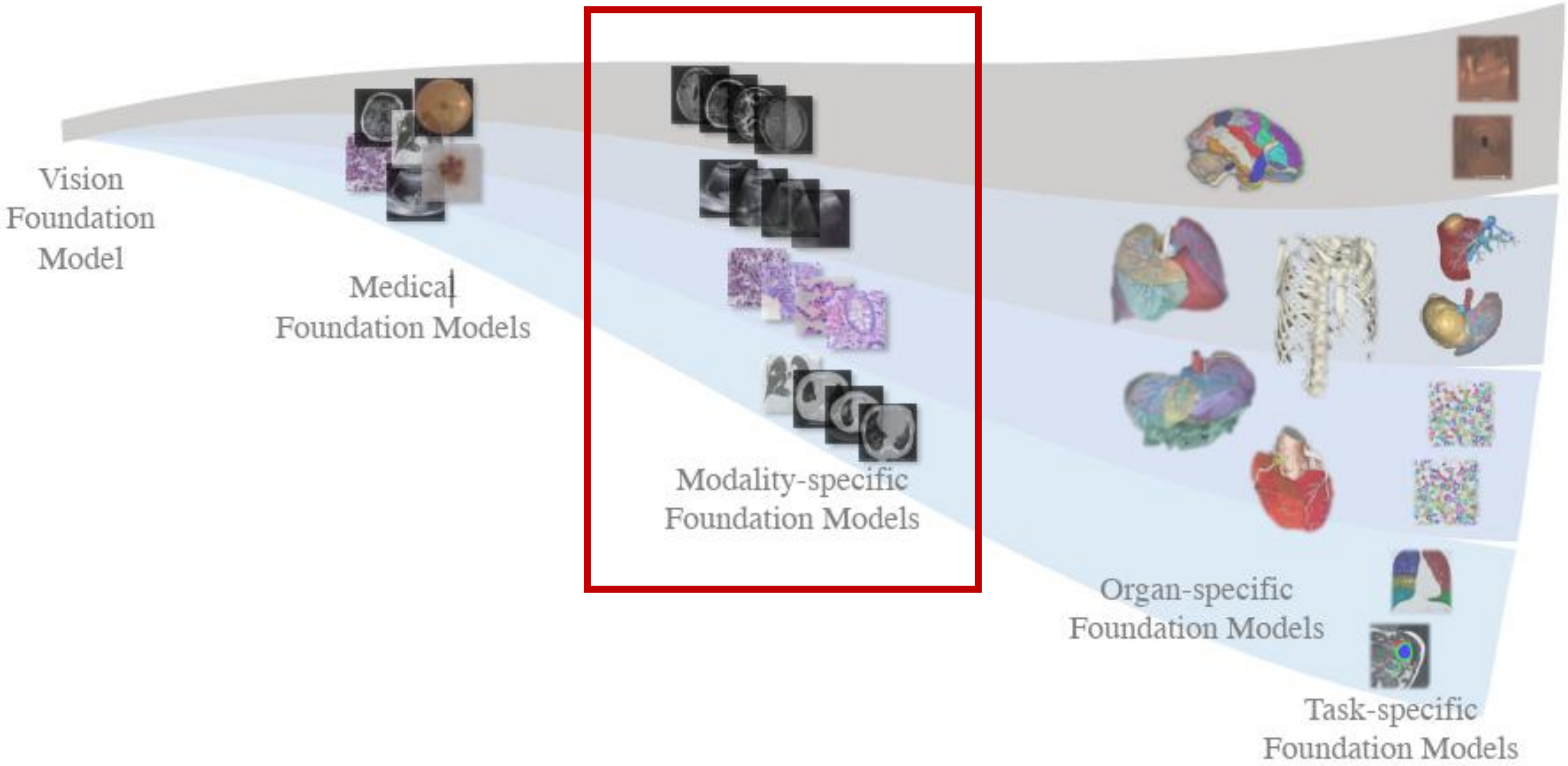
$$f_\theta(\mathbf{x}) = \mathbf{v}$$

How well generalist models transfer to specialized fundus image classification tasks?

(a) *Zero-shot*		MESSIDOR	FIVES	REFUGE	20x3	$ODIR_{200x3}$	MMAC	**Avg.**
CLIP	ViT-B/32	0.200	0.256	0.433	0.333	0.480	0.183	0.314
BiomedCLIP	ViT-B/16	0.207	0.415	0.624	0.617	0.583	0.274	0.453
(b) *Linear Probing*								
ImageNet	RN50	0.424	0.741	0.733	0.983	0.887	0.631	0.733
CLIP	ViT-B/32	0.491	0.800	0.720	0.950	0.917	0.642	0.753
BiomedCLIP	ViT-B/16	0.433	0.654	0.776	0.866	0.883	0.678	0.715

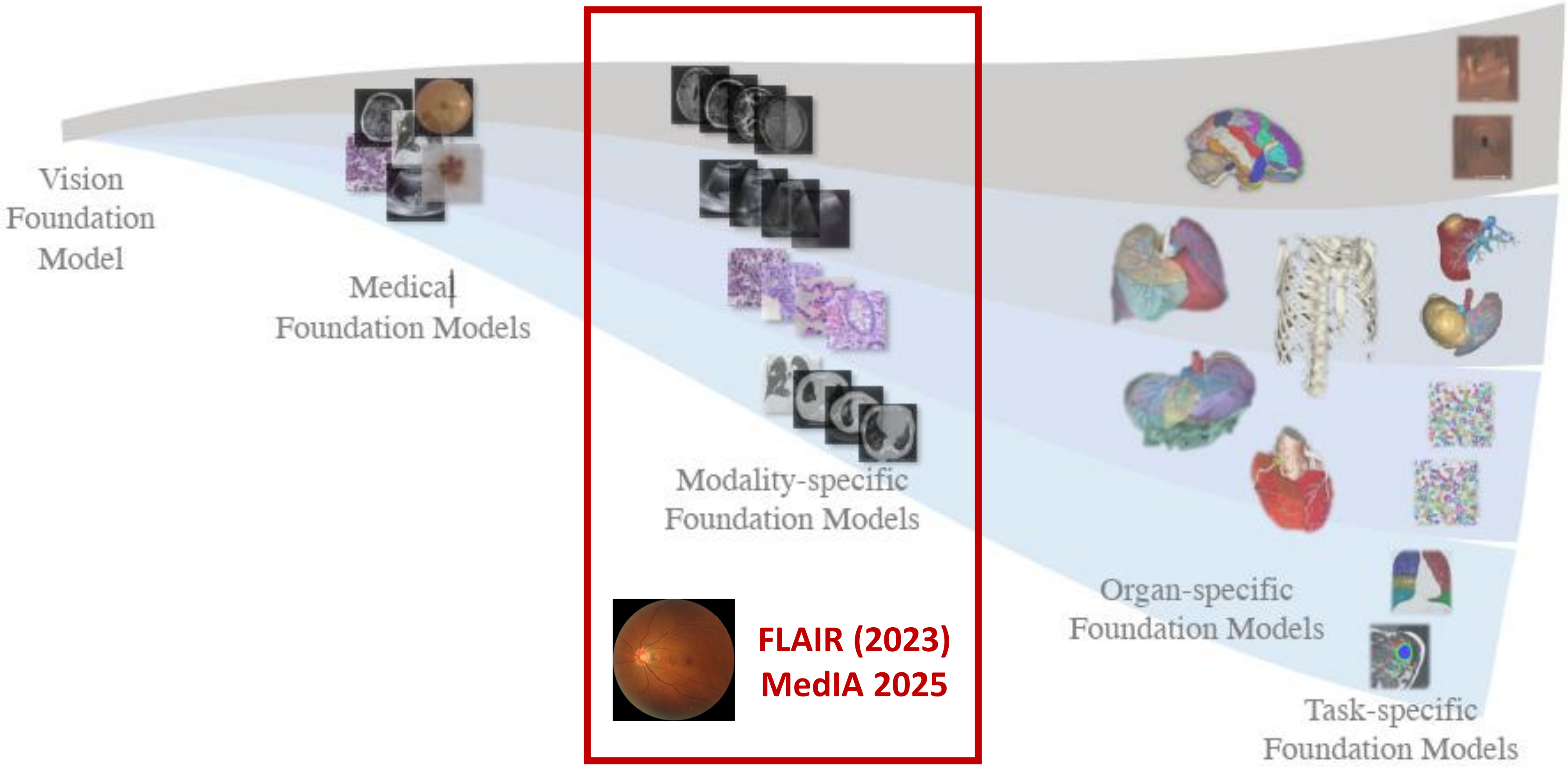
A Foundation Language-Image Model of the Retina (FLAIR): Encoding Expert Knowledge in Text Supervision, Medical Image Analysis (2025)

In search of a fundus image foundation model



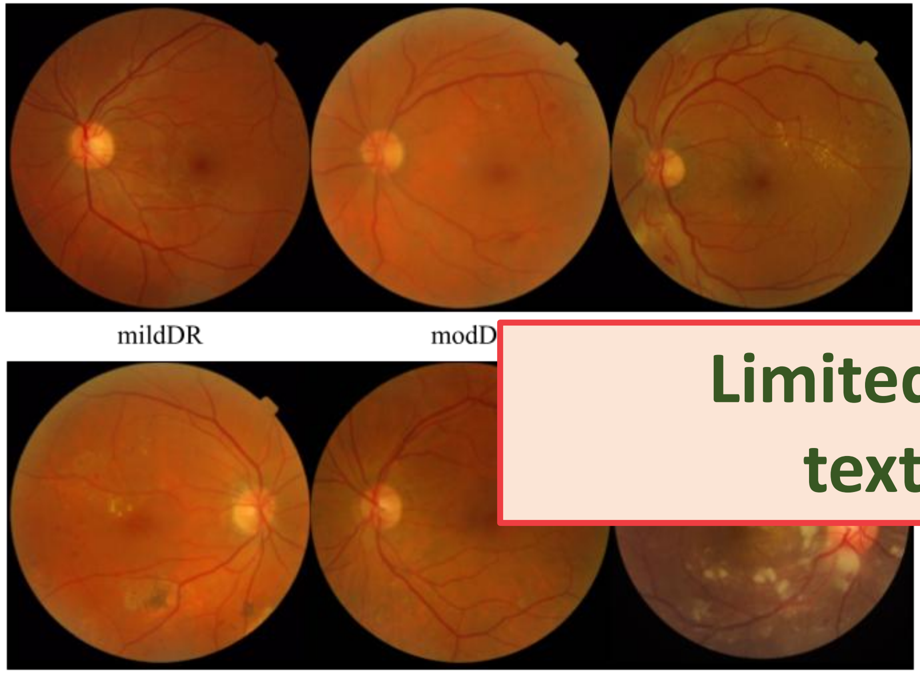
On the Challenges and Perspectives of Foundation Models for Medical Image Analysis, Medical Image Analysis (2024)

In search of a fundus image foundation model



On the Challenges and Perspectives of Foundation Models for Medical Image Analysis, Medical Image Analysis (2024)

Building an assembly dataset for pre-training



Limited datasets with text supervision

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.DEN (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
		28	DR, MA.	Categorical
		590	noDR, mildDR, modDR, sevDR, prolDR.	Categorical
		200	G, N, CME, neovARMD, geoARMD, acCSR, chCSR.	Categorical
		89	IrMA, neoV, ReSD, hEX, HE, sEX, MA.	Categorical
		110	noCAT, Dis.	Categorical
		100	N, G.	Categorical
		463	EX, MA.,	Categorical
		020	G, N.	Categorical
		169	EX, CWS, DN.	Categorical
		81	N, G, DR, noisy.	Categorical
		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		

Open-Access Datasets

A Foundation Language-Image Model of the Retina (FLAIR): Encoding Expert Knowledge in Text Supervision, Medical Image Analysis (2025)

Enhancing textual alingment trough expert-knowledge-driven descriptions

“moderate diabetic retinopathy”

“contains few microaneurysms”

“diabetic macular edema”

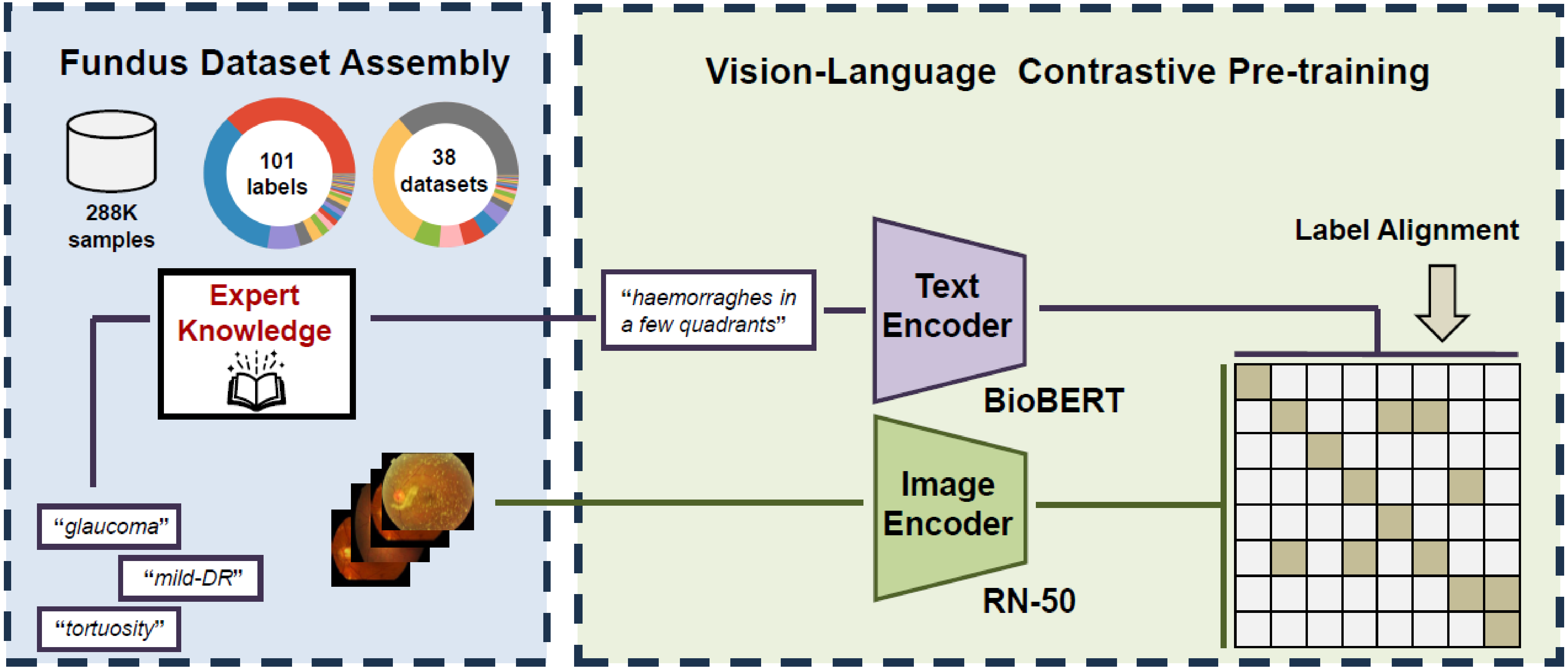
“exudates near the macula center”

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”

Expert Knowledge Dictionary

A Foundation Language-Image Model of the Retina (FLAIR): Encoding Expert Knowledge in Text Supervision, Medical Image Analysis (2025)

Image-Label-Text alignment

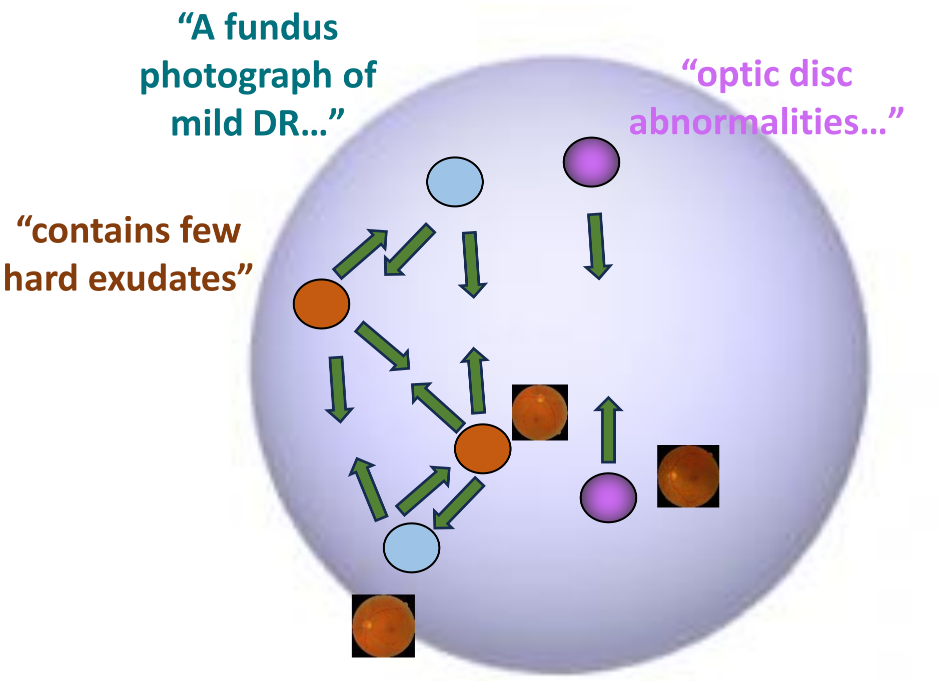


A Foundation Language-Image Model of the Retina (FLAIR): Encoding Expert Knowledge in Text Supervision, Medical Image Analysis (2025)

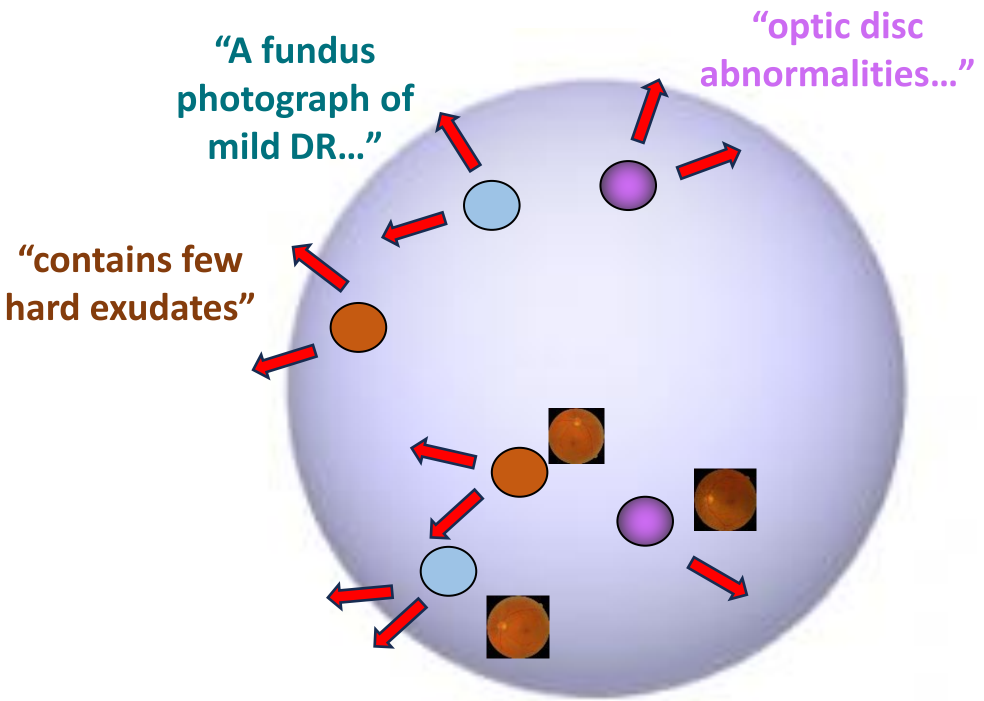
Image-Label-Text alignment

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

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



Approach points with same labels



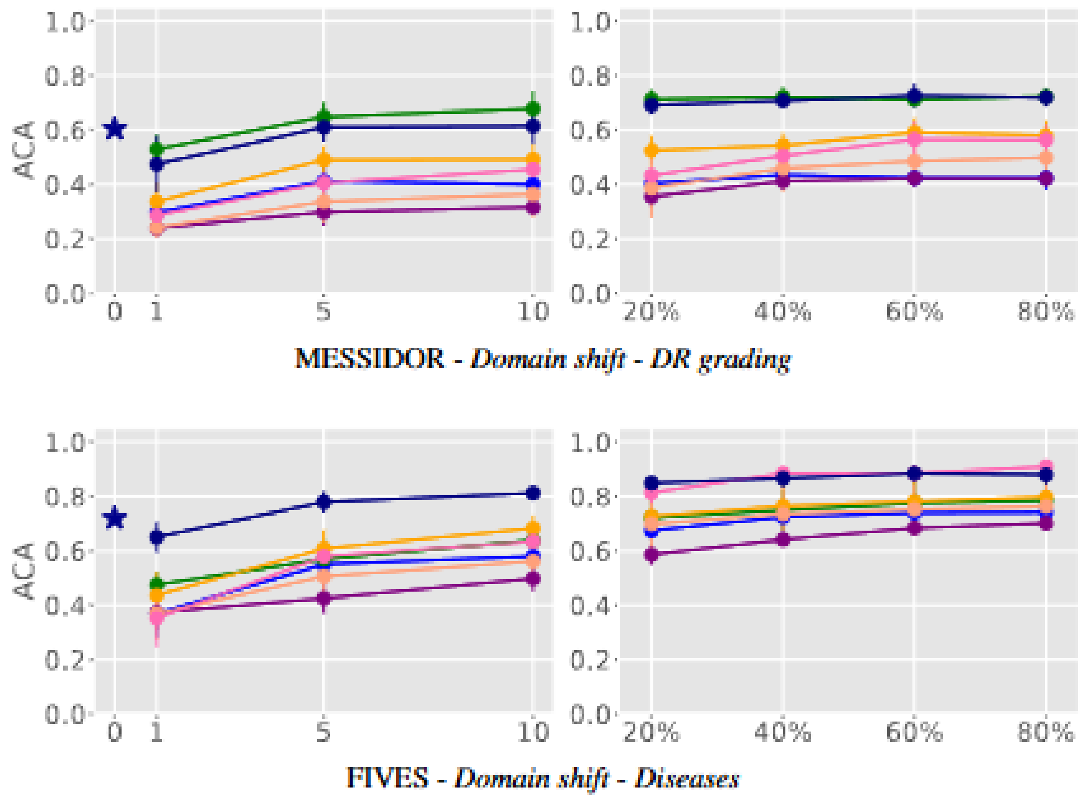
Push away points with different labels

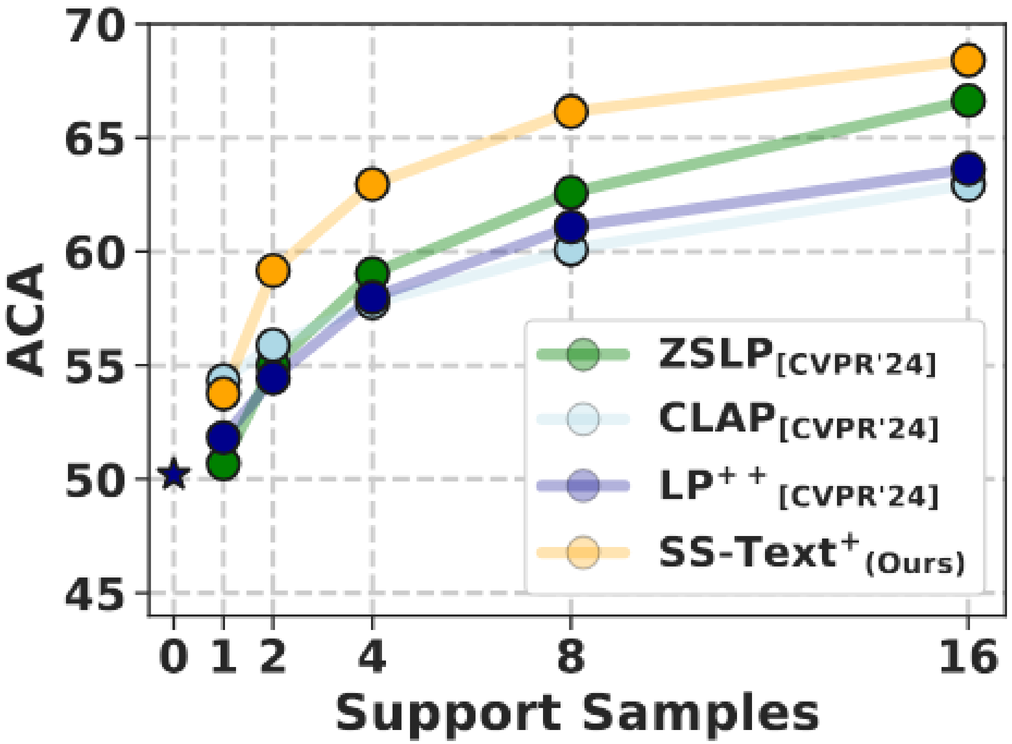
Zero-shot and Linear probing performance

(a) *Zero-shot*		MESSIDOR	FIVES	REFUGE	20x3	$ODIR_{200x3}$	MMAC	**Avg.**
CLIP	ViT-B/32	0.200	0.256	0.433	0.333	0.480	0.183	0.314
BiomedCLIP	ViT-B/16	0.207	0.415	0.624	0.617	0.583	0.274	0.453
FLAIR	RN50	**0.604**	**0.735**	**0.883**	**0.983**	**0.667**	**0.400**	**0.712**
(b) *Linear Probing*								
ImageNet	RN50	0.424	0.741	0.733	0.983	0.887	0.631	0.733
CLIP	ViT-B/32	0.491	0.800	0.720	0.950	0.917	0.642	0.753
BiomedCLIP	ViT-B/16	0.433	0.654	0.776	0.866	0.883	0.678	0.715
RETFound	ViT-B/16	0.457	0.765	0.747	0.950	0.887	0.547	0.725
FLAIR	RN50	**0.719**	**0.879**	**0.843**	**1.000**	**0.935**	**0.740**	**0.852**

A Foundation Language-Image Model of the Retina (FLAIR): Encoding Expert Knowledge in Text Supervision, Medical Image Analysis (2025)

Opportunities: efficient few-shot transfer

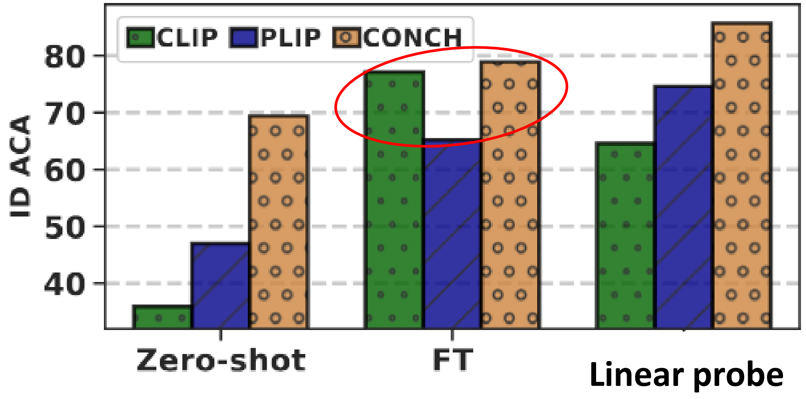


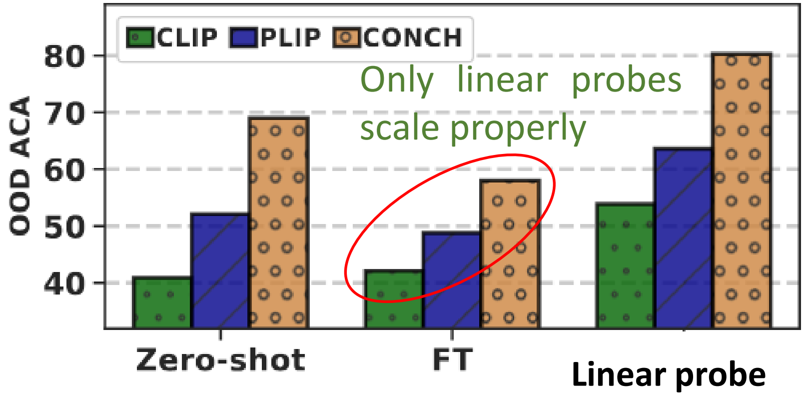


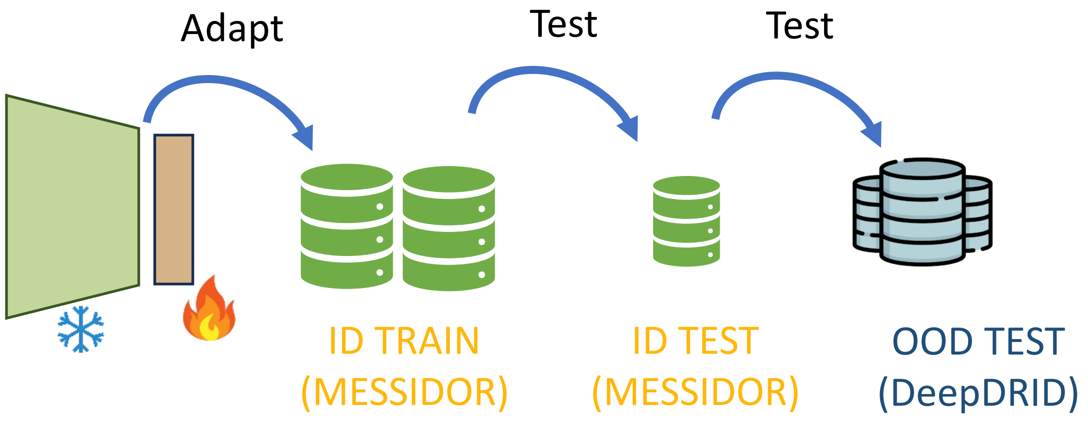
A Foundation Language-Image Model of the Retina (FLAIR): Encoding Expert Knowledge in Text Supervision, Medical Image Analysis (2025)
Few-Shot Adaptation of Medical Vision-Language Models, MICCAI (2024)
Few-Shot, Now for Real: Medical VLMs Adaptation without Balanced Sets or Validation, MICCAI (2025)

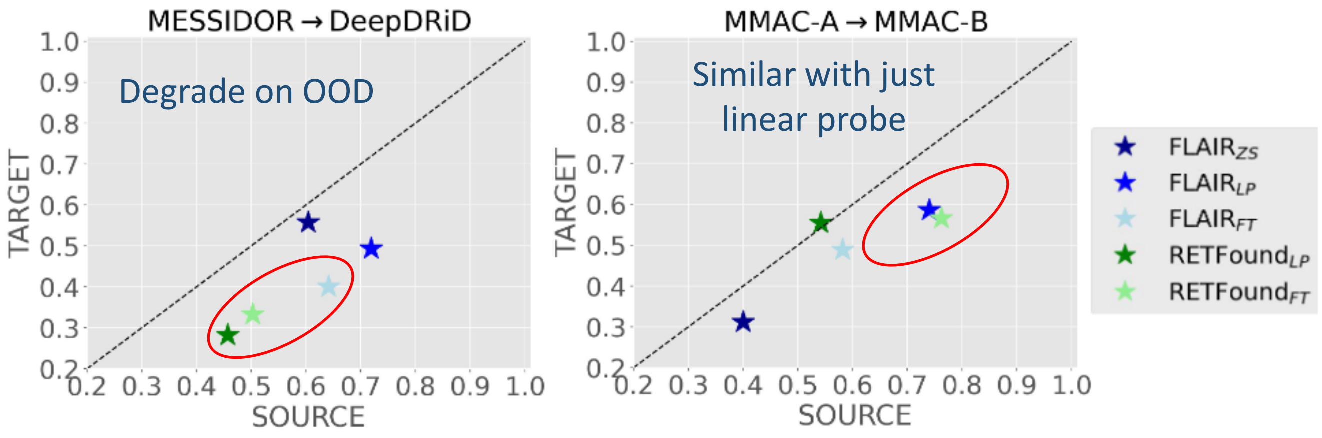
Opportunities: efficient domain generalization

For ID, full finetuning generalist models might produce good results









1. Limitations of full fine-tuning

2. Comparison with self-supervised models

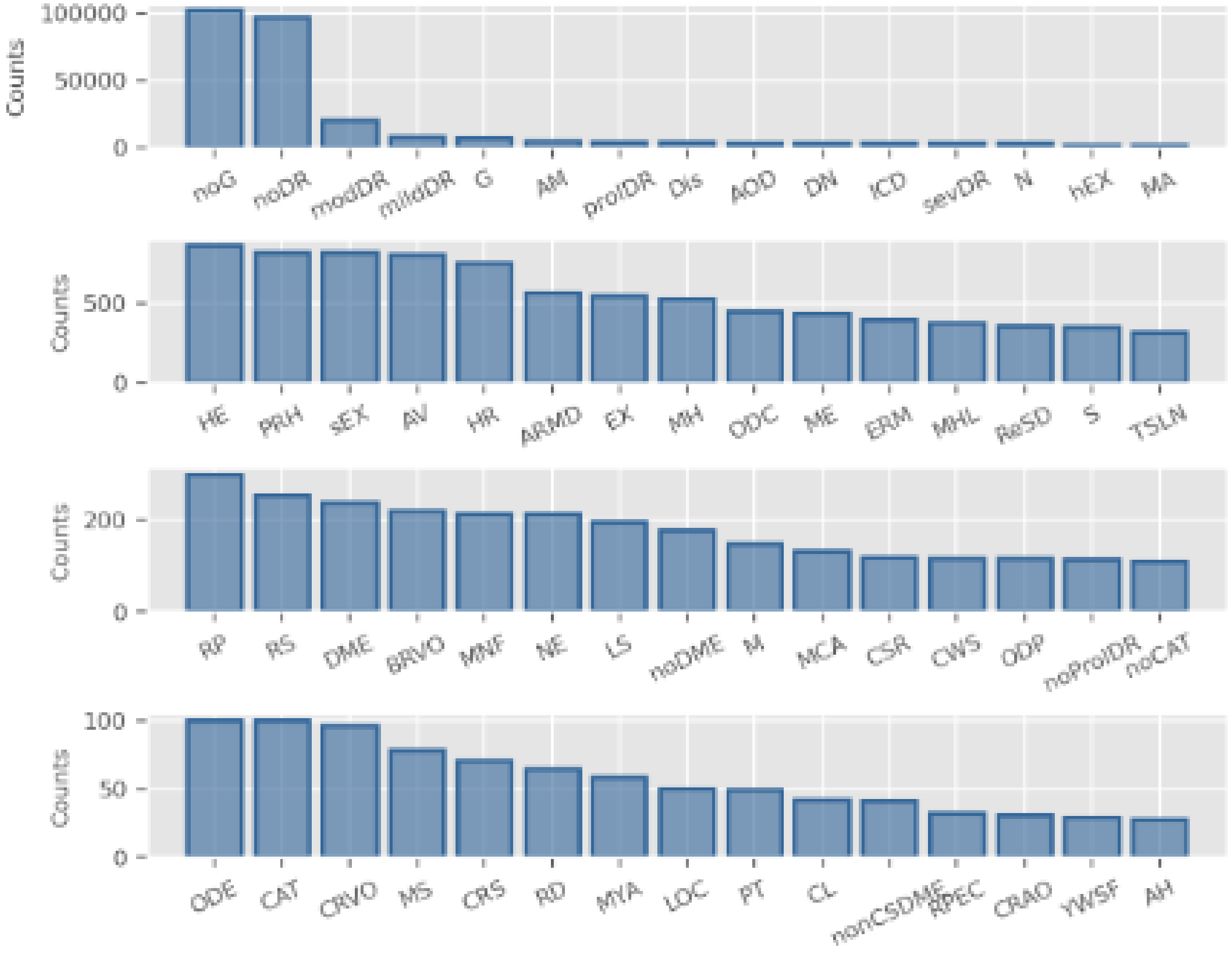
Robust Adaptation of Medical Vision-Language Models, On-going work.

Challenges: open-access datasets

Most pre-training data comes from very few categories:
DR grading, glaucoma detection, DME…

Pre-training concept frequencies shows a strong correlation with transferability.

Also, we have scarcity of textual data…

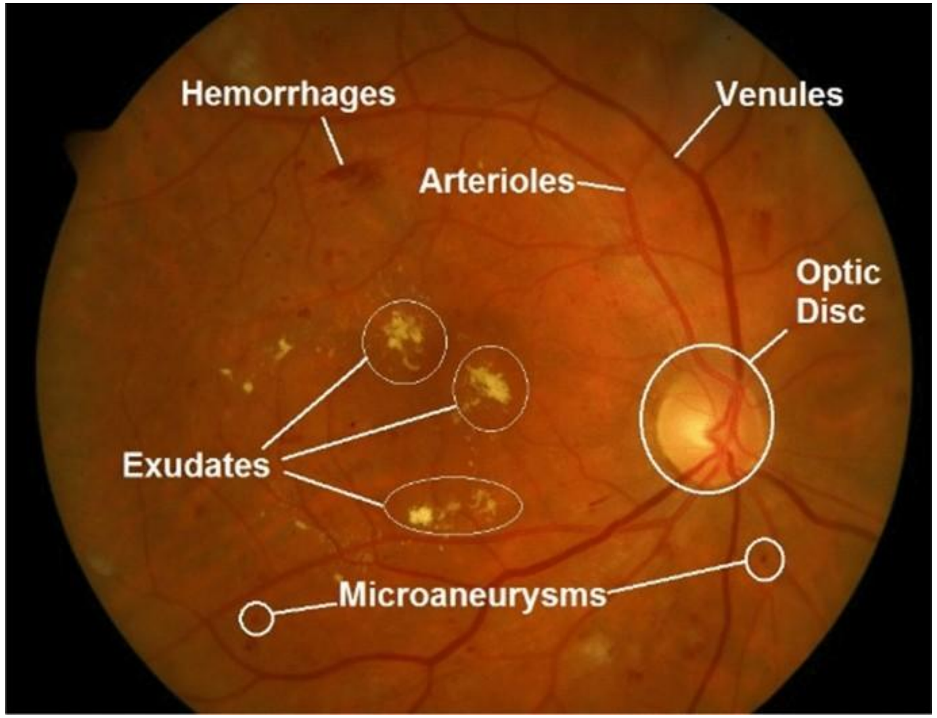


No “Zero-Shot” without Exponential Data , NeurIPS (2024)

Challenges: global + local information

Current pre-training strategies following CLIP leverage global embedding representations. However, critical findings in fundus images are local, sparsely located.

We need a better understanding of fine-grained patterns in the multi-modal space, e.g., relative location, size, etc.



```python
from PIL import Image
import numpy as np

# Import FLAIR
from flair import FLAIRModel

# Set model
model = FLAIRModel(from_checkpoint=True)

# Load image and set target categories
# (if the repo is not cloned, download the image and change the path!)

image = np.array(Image.open("./documents/sample_macular_hole.png"))
text = ["normal", "healthy", "macular edema", "diabetic retinopathy", "glaucoma", "macular hole",
        "lesion", "lesion in the macula"]

# Forward FLAIR model to compute similarities
probs, logits = model(image, text)

print("Image-Text similarities:")
print(logits.round(3)) # [[-0.32  -2.782  3.164  4.388  5.919  6.639  6.579 10.478]]
print("Probabilities:")
print(probs.round(3))  # [[0.      0.     0.001  0.002  0.01   0.02   0.019  0.948]]
```

★ **Try FLAIR!**

https://github.com/jusiro/FLAIR