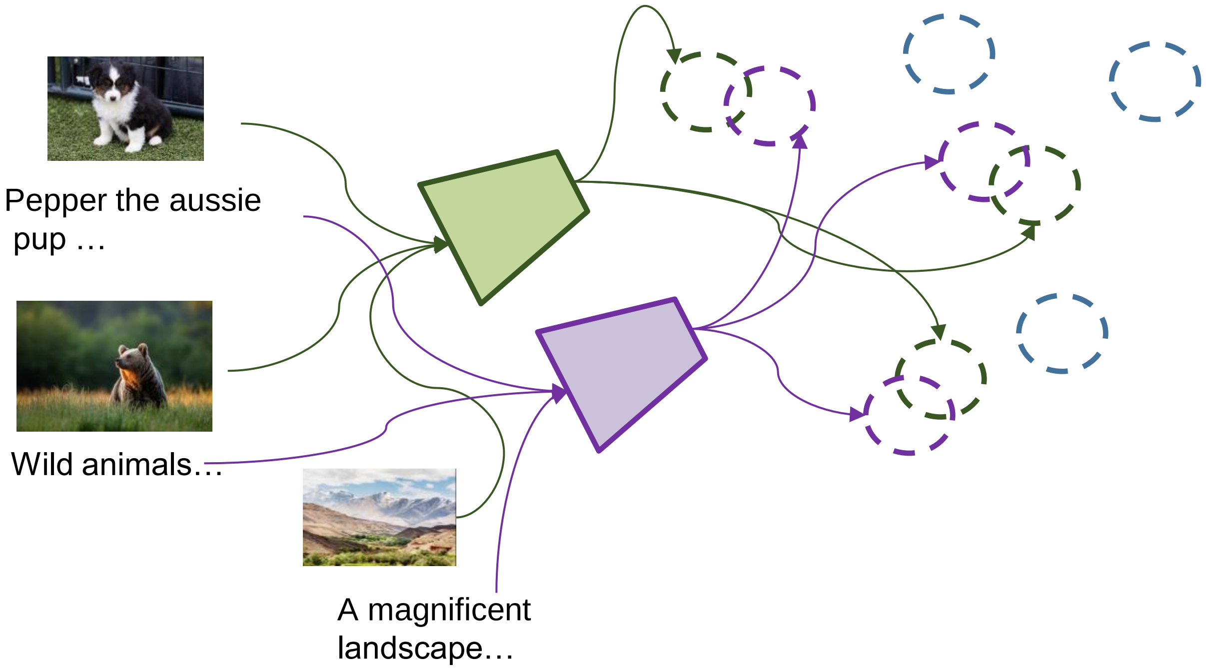
Few-shot Adaptation of Medical Vision-Language Models

Fereshteh Shakeri
Yunshi Huang
Julio Silva-Rodríguez
Houda Bahig
An Tang
Jose Dolz
Ismail Ben Ayed

Generalist Vision-Language Models (VLMs)

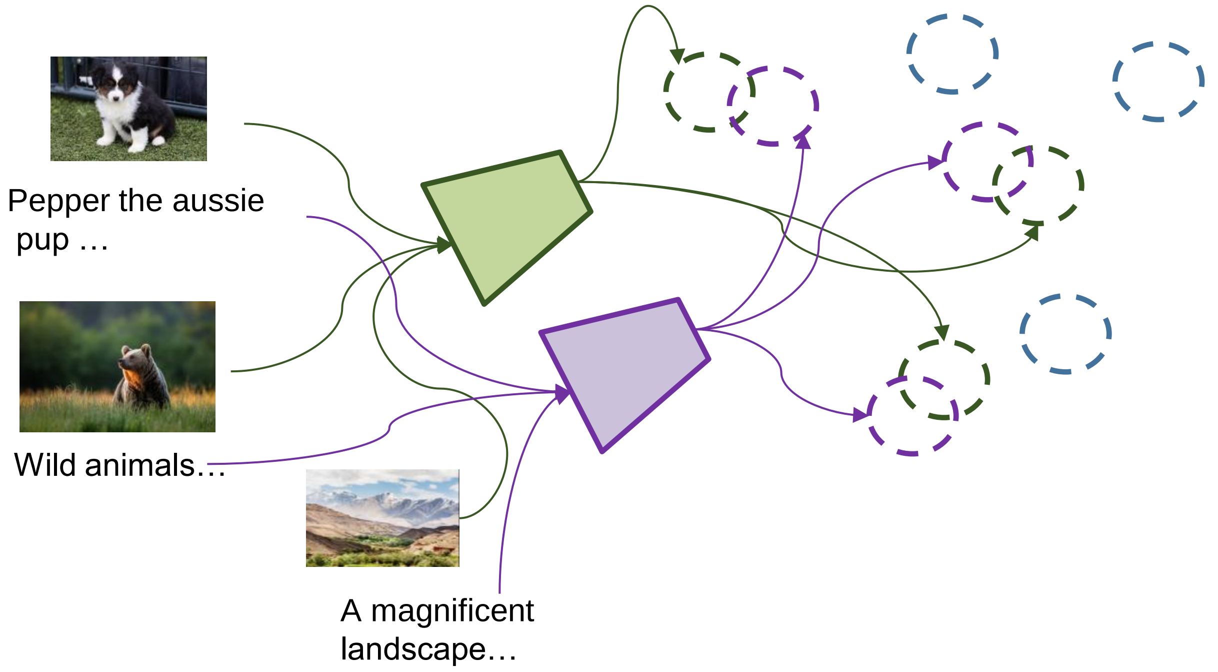
Unsupervised image-language pre-training



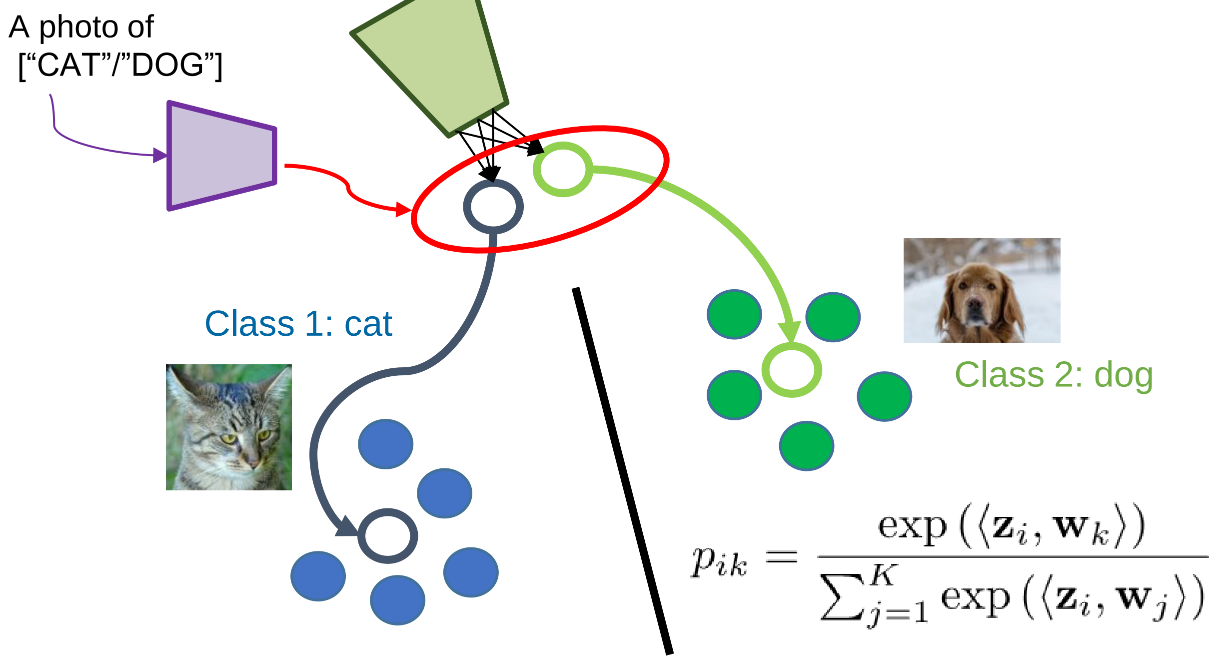
CLIP: Radford et al., Learning transferable visual models from natural language supervision, ICML 2021

Generalist Vision-Language Models (VLMs)

Unsupervised image-language pre-training



Zero-shot

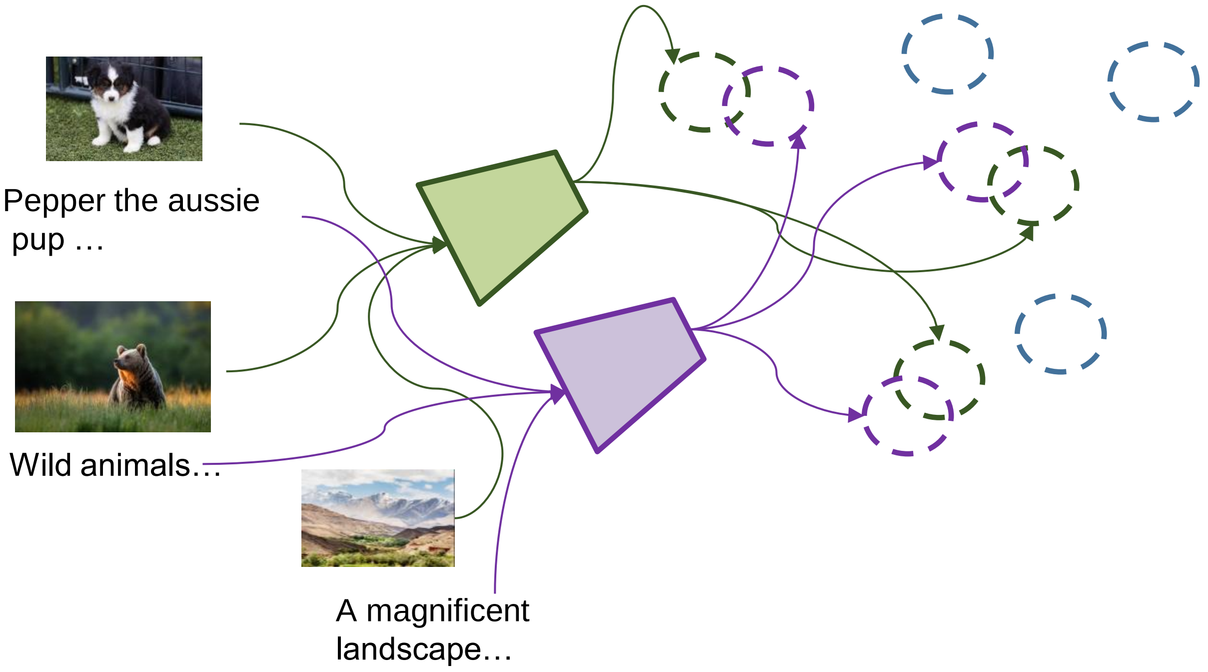


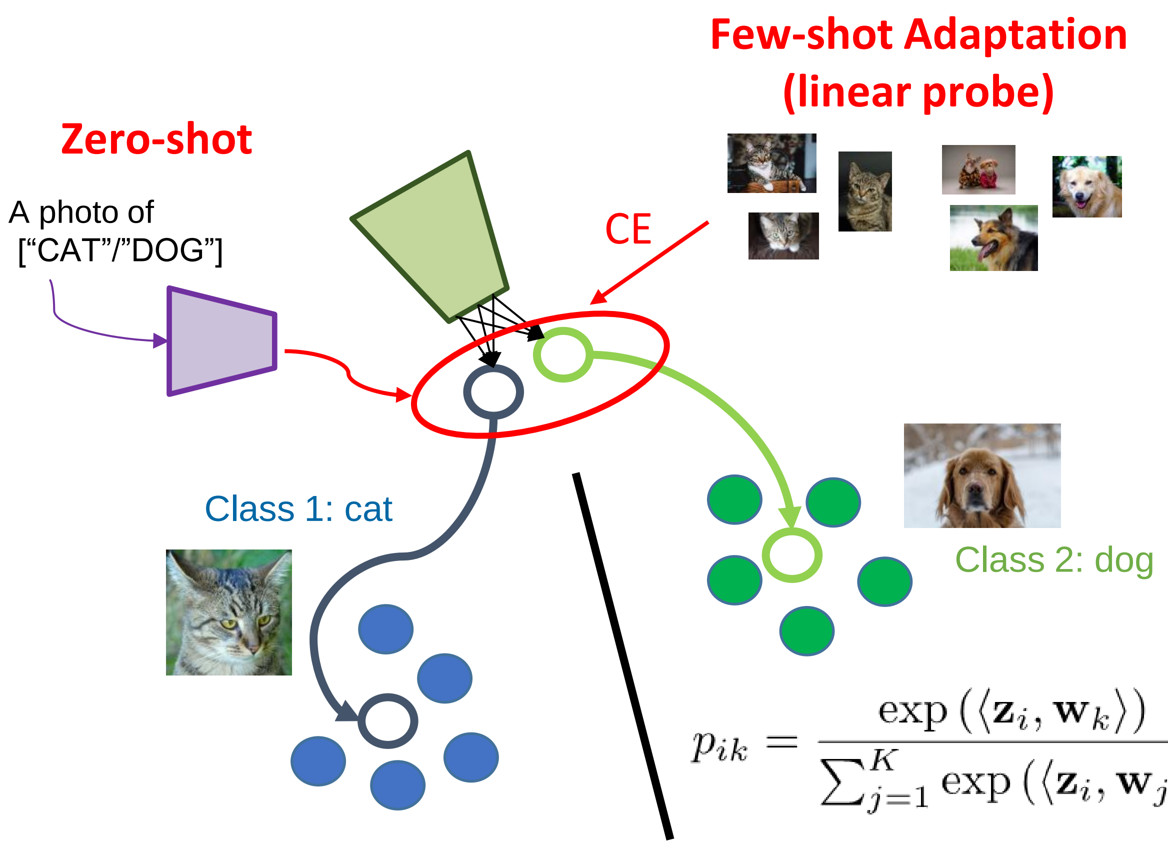
$$p_{ik} = \frac{\exp\left(\langle \mathbf{z}_i, \mathbf{w}_k \rangle\right)}{\sum_{j=1}^{K} \exp\left(\langle \mathbf{z}_i, \mathbf{w}_j \rangle\right)}$$

CLIP: Radford et al., Learning transferable visual models from natural language supervision, ICML 2021

Generalist Vision-Language Models (VLMs)

Unsupervised image-language pre-training



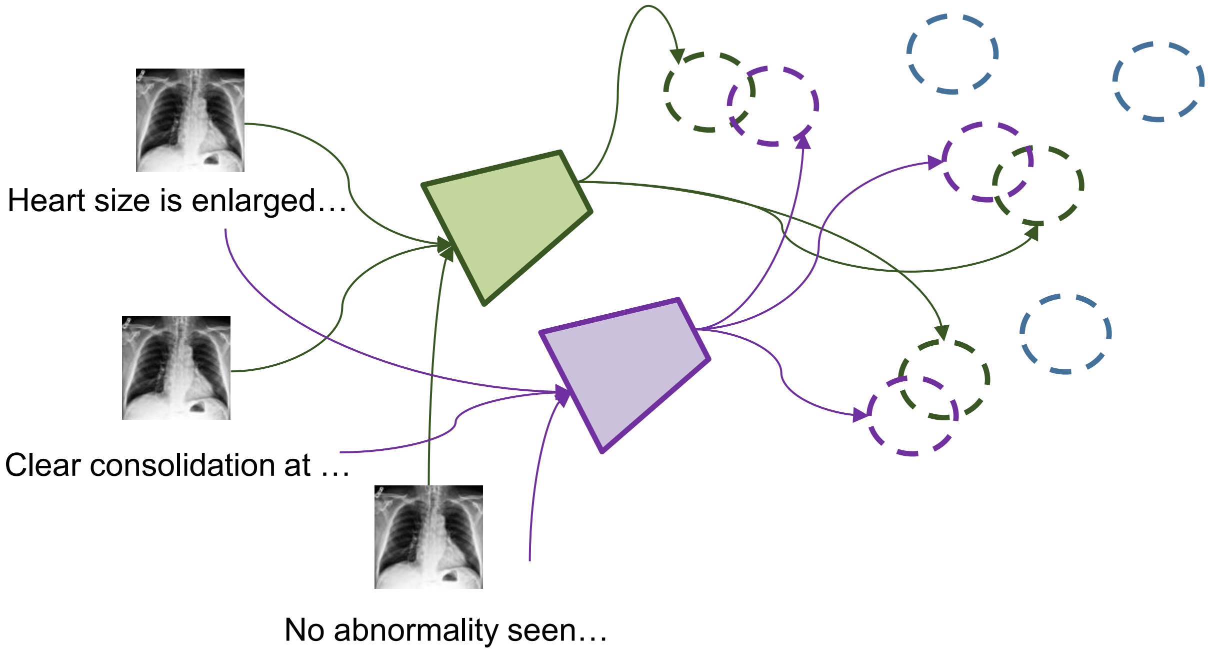


$$p_{ik} = \frac{\exp\left(\langle \mathbf{z}_i, \mathbf{w}_k \rangle\right)}{\sum_{j=1}^{K} \exp\left(\langle \mathbf{z}_i, \mathbf{w}_j \rangle\right)}$$

CLIP: Radford et al., Learning transferable visual models from natural language supervision, ICML 2021

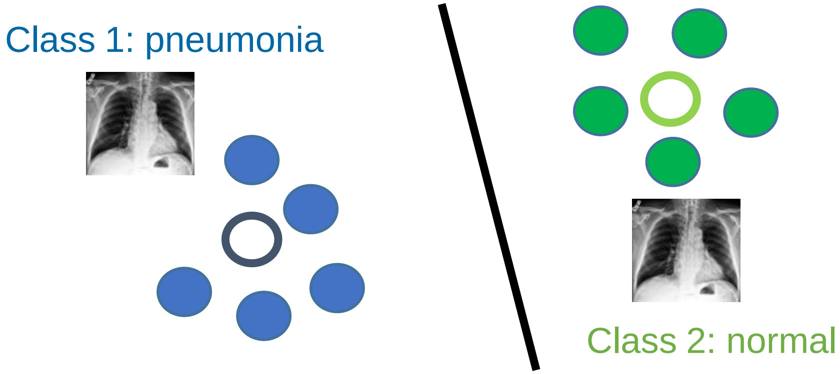
~~Generalist~~ Medical **(Specialized)** VLMs

Unsupervised image-language pre-training



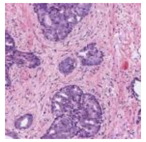
CONVIRT: Zhang et al., Medical Visual Representations from Paired Images and Text, MLHC 2022
MedCLIP: Wang et al., Contrastive Learning from Unpaired medical images and text, EMNLP 2022
Quilt-1M: Ikezogwo et al., One Million Image-Text Pairs for Histopathology, NeurIPS 2023
FLAIR: Silva-Rodríguez et al., A Foundation Language-Image Model of the Retina, MedIA 2024
…

Zero-shot / Few-shot Adaptation

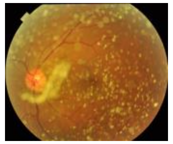


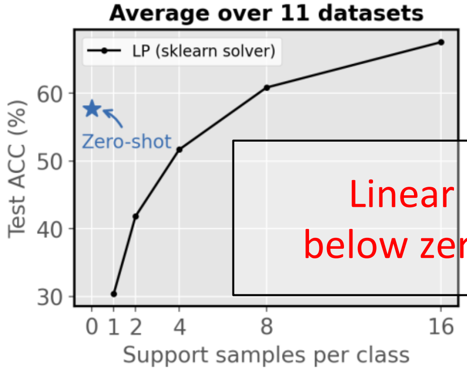
MedCLIP

Quilt-1M



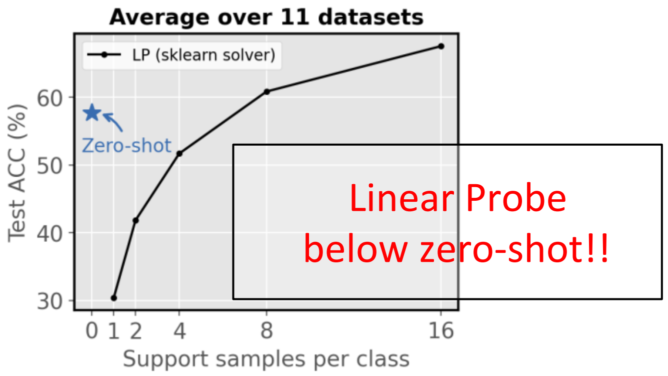
FLAIR

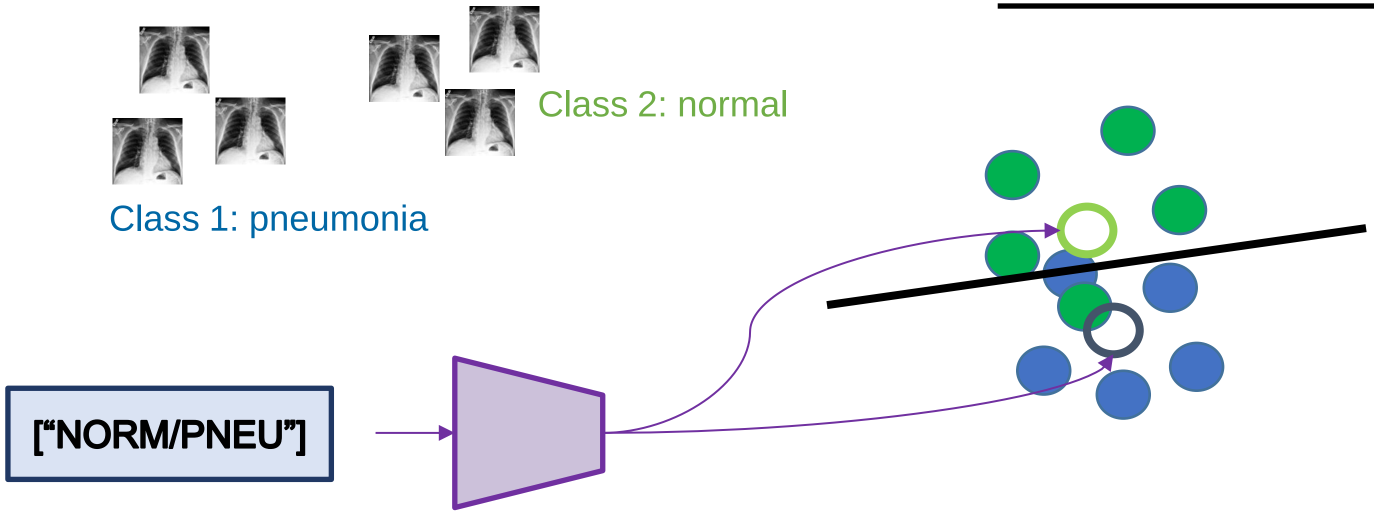




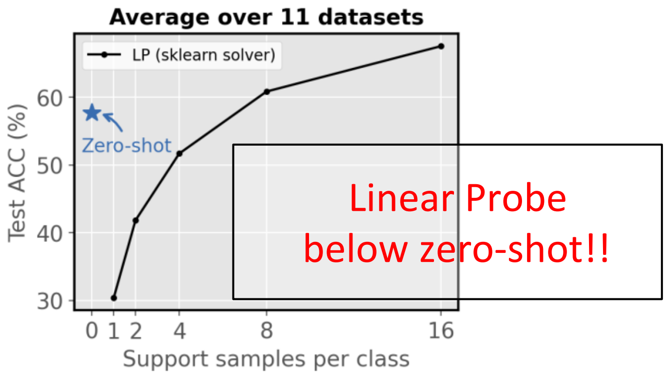
Linear Probe
below zero-shot!!

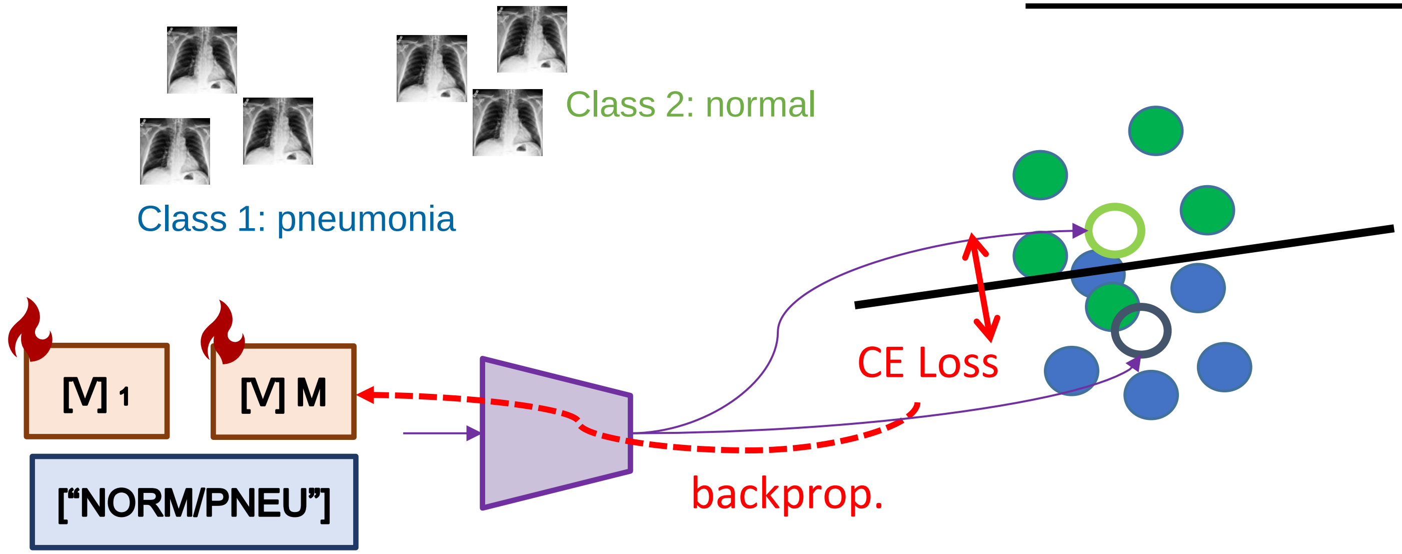
(Popular) Prompt Learning

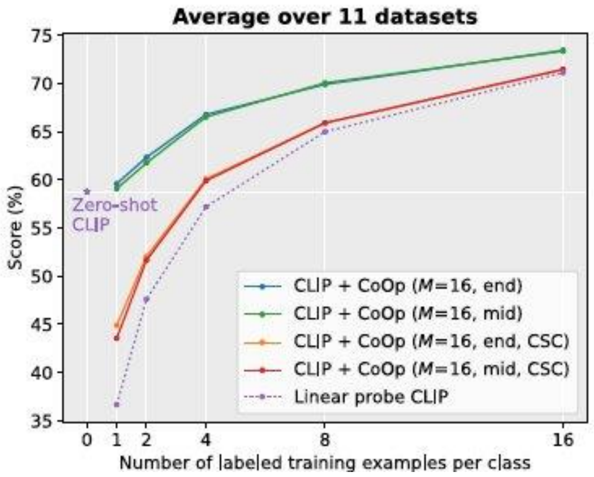




(Popular) Prompt Learning







CoOp: Zhou et al., Learning to Prompt for Vision-Language Models, IJCV 2022[~1900 citations]
CoCoOp: Zhou et al., Conditional Prompt Learning for Vision-Language Models, CVPR 2022[~1200 citations]
KgCoOp: Yao et al., Prompt Tuning with Knowledge-guided Context Optimization, CVPR 2023[~122 citations]

Towards better **black-box Adapters**: LP+text

"Weak Baseline" Linear Probe

Features

Learned
weight

$$p_{ik} = \frac{\exp\left(\langle \mathbf{z}_i, \mathbf{w}_k \rangle\right)}{\sum_{j=1}^{K} \exp\left(\langle \mathbf{z}_i, \mathbf{w}_j \rangle\right)}$$

(only vision)
Neglects any text knowledge

LP++: Huang et al., A Surprisingly Strong Linear Probe for Few-Shot CLIP, CVPR 2024

Towards better **black-box Adapters**: LP+text

"Weak Baseline" Linear Probe

Features

Learned weight

$$p_{ik} = \frac{\exp\left(\langle \mathbf{z}_i, \mathbf{w}_k \rangle\right)}{\sum_{j=1}^{K} \exp\left(\langle \mathbf{z}_i, \mathbf{w}_j \rangle\right)}$$

(only vision)
Neglects any text knowledge

Text- Informed Linear Probe

Text (zero-shot) prototype

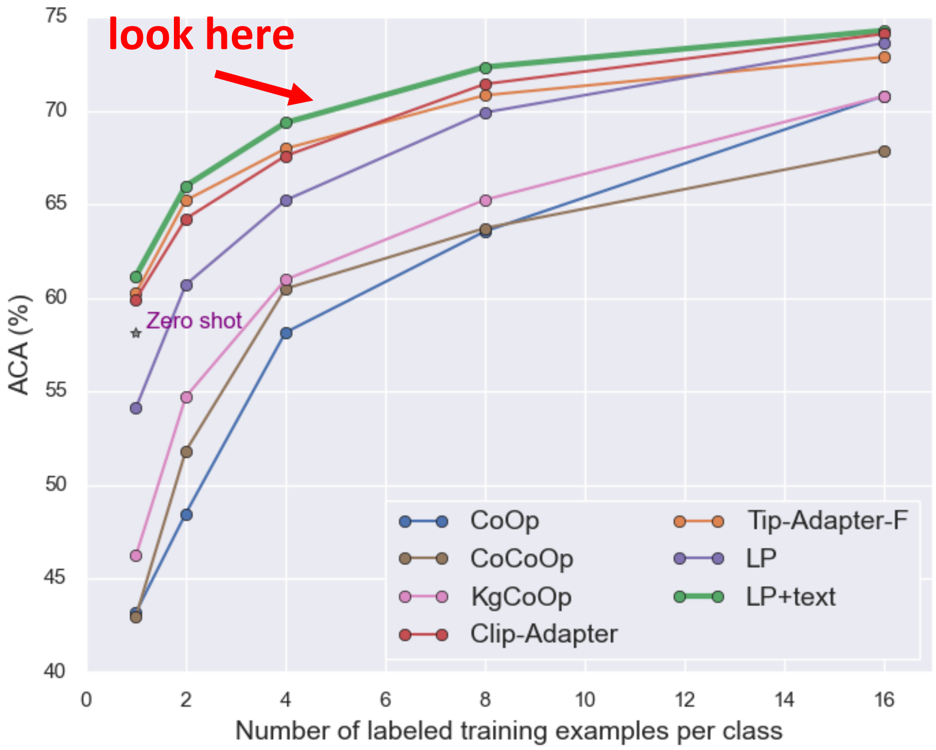
$$p_{ik} = \frac{\exp\left(\langle \mathbf{z}_i, \mathbf{w}_k + \alpha_k \boldsymbol{t}_k \rangle\right)}{\sum_{j=1}^{K} \exp\left(\langle \mathbf{z}_i, \mathbf{w}_j + \alpha_j \boldsymbol{t}_j \rangle\right)}$$

Trainable image-text blending weight

LP++: Huang et al., A Surprisingly Strong Linear Probe for Few-Shot CLIP, CVPR 2024

Results

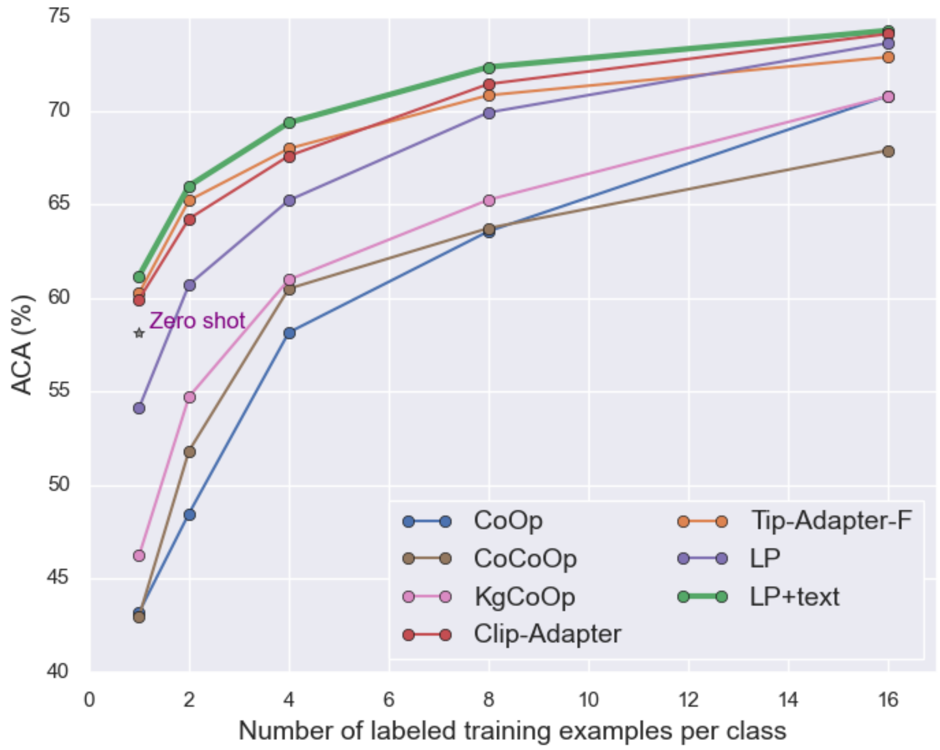
3 modalities / 9 datasets



LP+text is <u>competitive</u>

Results

3 modalities / 9 datasets



LP+text is <u>competitive</u>

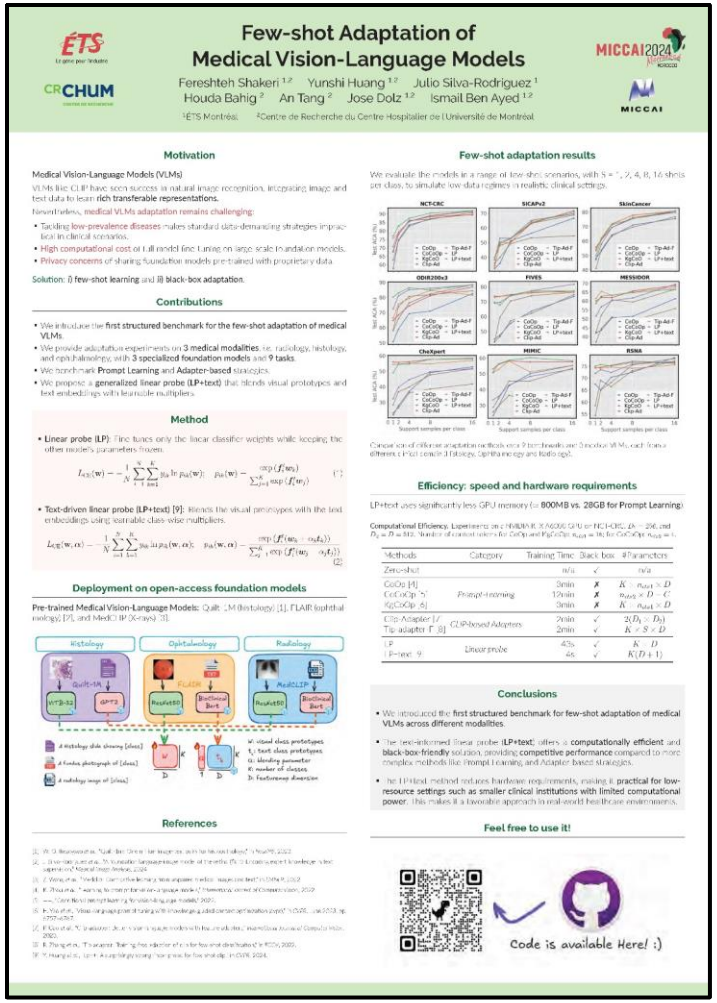
LP+text is <u>extremely efficient</u>!

→ Adaptation in a **matter of seconds**

→ Trainable on **commodity GPUs (MBs)**

→ **Black-box adaptation**

Methods	Category	Training Time	Black-box	#Parameters
Zero-shot		n/a	✓	n/a
CoOp [4]	*Prompt-Learning*	3min	✗	$K \times n_{ctx1} \times D$
CoCoOp [5]		12min	✗	$n_{ctx2} \times D + C$
KgCoOp [6]		3min	✗	$K \times n_{ctx1} \times D$
Clip-Adapter [7]	*CLIP-based Adapters*	2min	✓	$2(D_1 \times D_2)$
Tip-adapter-F [8]		2min	✓	$K \times S \times D$
LP	*Linear probe*	43s	✓	$K \times D$
★ LP+text [9]		4s	✓	$K(D+1)$



TUESDAY-PM

POSTER 096

Any questions?

Open Benchmark!