Few-shot Adaptation of Medical Vision-Language Models

Fereshteh Shakeri [1,2] Yunshi Huang [1,2] Julio Silva-Rodríguez [1]
Houda Bahig [2] An Tang [2] Jose Dolz [1,2] Ismail Ben Ayed [1,2]

[1]ÉTS Montréal [2]Centre de Recherche du Centre Hospitalier de l'Université de Montréal

Motivation

Medical Vision-Language Models (VLMs)

VLMs like CLIP have seen success in natural image recognition, integrating image and text data to learn **rich transferable representations.**

Nevertheless, medical VLMs adaptation remains challenging:

- Tackling low-prevalence diseases makes standard data-demanding strategies impractical in clinical scenarios.
- High computational cost of full model fine-tuning on large-scale foundation models.
- Privacy concerns of sharing foundation models pre-trained with proprietary data.

Solution: *i)* few-shot learning and *ii)* black-box adaptation.

Contributions

- We introduce the **first structured benchmark for the few-shot adaptation of medical VLMs.**
- We provide adaptation experiments on **3 medical modalities**, *i.e.* radiology, histology, and ophthalmology, with **3 specialized foundation models** and **9 tasks**.
- We benchmark **Prompt Learning** and **Adapter-based** strategies.
- We propose a **generalized linear probe (LP+text)** that blends visual prototypes and text embeddings with learnable multipliers.

Method

- **Linear probe (LP)**: Fine-tunes only the linear-classifier weights while keeping the other model's parameters frozen.

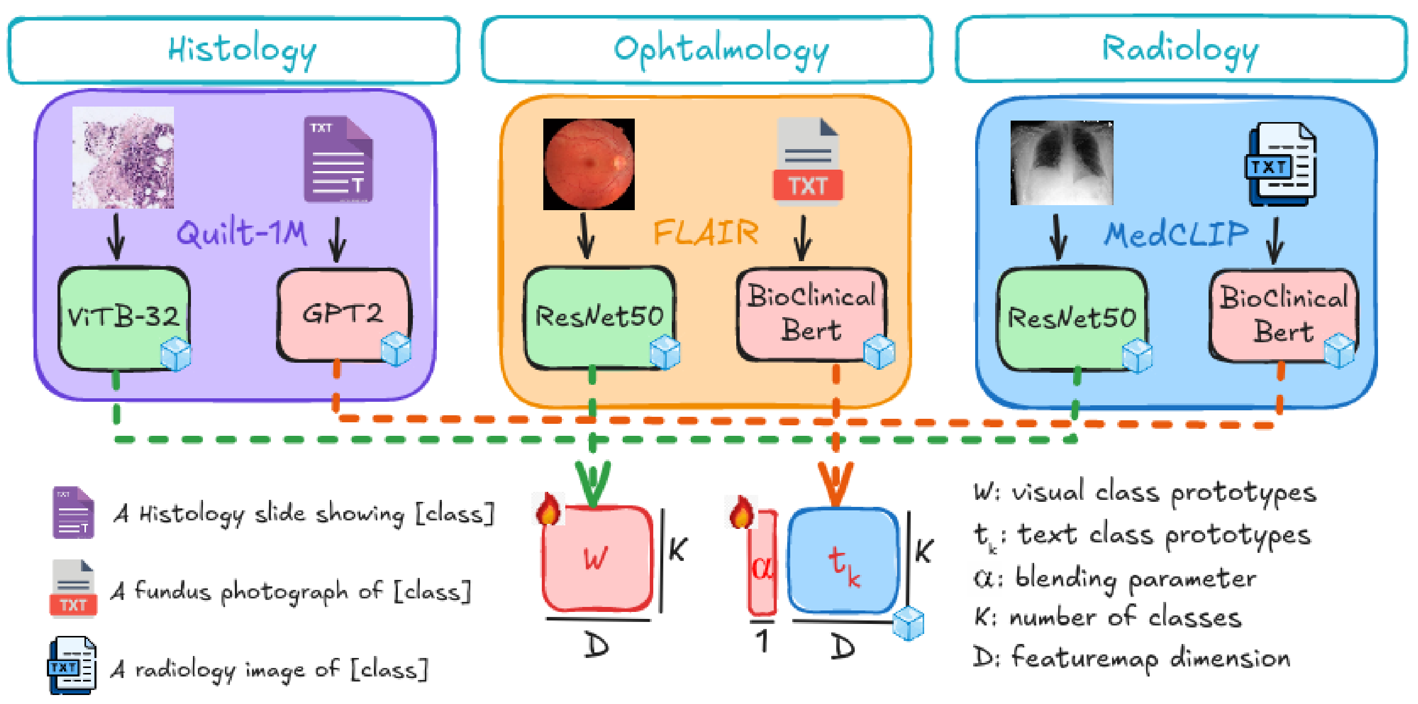
$$L_{\mathrm{CE}}(\mathbf{w}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\ln p_{ik}(\mathbf{w}); \quad p_{ik}(\mathbf{w}) = \frac{\exp\left(\boldsymbol{f}_i^t \boldsymbol{w}_k\right)}{\sum_{j=1}^{K}\exp\left(\boldsymbol{f}_i^t \boldsymbol{w}_j\right)} \tag{1}$$

- **Text-driven linear probe (LP+text) [9]:** Blends the visual prototypes with the text embeddings using learnable class-wise multipliers.

$$L_{\mathrm{CE}}(\mathbf{w},\boldsymbol{\alpha}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\ln p_{ik}(\mathbf{w},\boldsymbol{\alpha}); \quad p_{ik}(\mathbf{w},\boldsymbol{\alpha}) = \frac{\exp\left(\boldsymbol{f}_i^t(\boldsymbol{w}_k+\alpha_k\boldsymbol{t}_k)\right)}{\sum_{j=1}^{K}\exp\left(\boldsymbol{f}_i^t(\boldsymbol{w}_j+\alpha_j\boldsymbol{t}_j)\right)} \tag{2}$$

Deployment on open-access foundation models

Pre-trained Medical Vision-Language Models: Quilt-1M (histology) [1], FLAIR (ophthalmology) [2], and MedCLIP (X-rays) [3].



References

[1] W. O. Ikezogwo *et al.*, "Quilt-1m: One million image-text pairs for histopathology," in *NeurIPS*, 2023.

[2] J. Silva-Rodriguez *et al.*, "A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision," *Medical Image Analysis*, 2024.

[3] Z. Wang *et al.*, "Medclip: Contrastive learning from unpaired medical images and text," in *EMNLP*, 2022.

[4] K. Zhou *et al.*, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, 2022.

[5] ——, "Conditional prompt learning for vision-language models," 2022.

[6] H. Yao *et al.*, "Visual-language prompt tuning with knowledge-guided context optimization (cvpr)," in *CVPR*, June 2023, pp. 6757–6767.

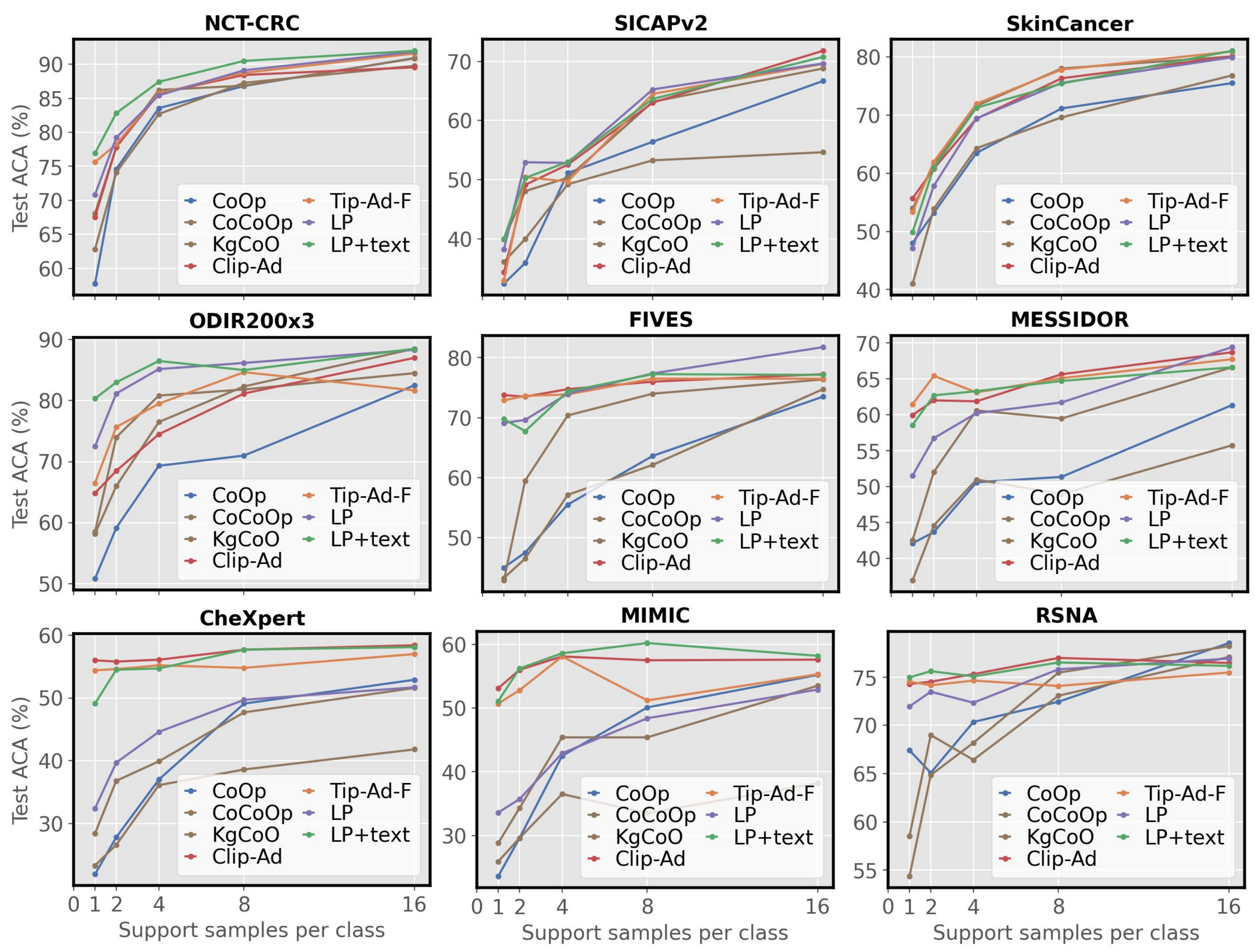
[7] P. Gao *et al.*, "Clip-adapter: Better vision-language models with feature adapters," *International Journal of Computer Vision*, 2023.

[8] R. Zhang *et al.*, "Tip-adapter: Training-free adaption of clip for few-shot classification," in *ECCV*, 2022.

[9] Y. Huang *et al.*, "Lp++: A surprisingly strong linear probe for few-shot clip," in *CVPR*, 2024.

Few-shot adaptation results

We evaluate the models in a range of few-shot scenarios, with S = 1, 2, 4, 8, 16 shots per class, to simulate low-data regimes in realistic clinical settings.



Comparison of different adaptation methods over 9 benchmarks and 3 medical VLMs, each from a different clinical domain (Histology, Ophthalmology and Radiology).

Efficiency: speed and hardware requirements

LP+text uses significantly less GPU memory (≃ **800MB vs. 28GB for Prompt Learning**).

Computational Efficiency. Experiments on a NVIDIA RTX A6000 GPU on NCT-CRC. $D_1 = 256$, and $D_2 = D = 512$. Number of context tokens for CoOp and KgCoOp: $n_{ctx1} = 16$; for CoCoOp: $n_{ctx2} = 4$.

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]	*Prompt-Learning*	12min	✗	$n_{ctx2} \times D + C$
KgCoOp [6]	*Prompt-Learning*	3min	✗	$K \times n_{ctx1} \times D$
Clip-Adapter [7]	*CLIP-based Adapters*	2min	✓	$2(D_1 \times D_2)$
Tip-adapter-F [8]	*CLIP-based Adapters*	2min	✓	$K \times S \times D$
LP	*Linear probe*	43s	✓	$K \times D$
LP+text [9]	*Linear probe*	4s	✓	$K(D+1)$

Conclusions

- We introduced the **first structured benchmark for few-shot adaptation of medical VLMs across different modalities.**
- The text-informed linear probe (**LP+text**) offers a **computationally efficient** and **black-box-friendly** solution, providing **competitive performance** compared to more complex methods like Prompt Learning and Adapter-based strategies.
- The LP+text method reduces hardware requirements, making it **practical for low-resource settings such as smaller clinical institutions with limited computational power.** This makes it a favorable approach in real-world healthcare environments.

Feel free to use it!

Code is available Here! :)