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], and Ismail Ben Ayed[1,2]

[1]ÉTS Montreal
[2]Centre de Recherche du Centre Hospitalier de l'Université de Montréal (CRCHUM)

Abstract. Integrating image and text data through multi-modal learning has emerged as a new approach in medical imaging research, following its successful deployment in computer vision. While considerable efforts have been dedicated to establishing medical foundation models and their zero-shot transfer to downstream tasks, the popular few-shot setting remains relatively unexplored. Following on from the currently strong emergence of this setting in computer vision, we introduce the first structured benchmark for adapting medical vision-language models (VLMs) in a strict few-shot regime and investigate various adaptation strategies commonly used in the context of natural images. Furthermore, we evaluate a simple generalization of the linear-probe adaptation baseline, which seeks an optimal blending of the visual prototypes and text embeddings via learnable class-wise multipliers. Surprisingly, such a text-informed linear probe yields competitive performances in comparison to convoluted prompt-learning and adapter-based strategies, while running considerably faster and accommodating the black-box setting. Our extensive experiments span three different medical modalities and specialized foundation models, nine downstream tasks, and several state-of-the-art few-shot adaptation methods. We made our benchmark and code publicly available to trigger further developments in this emergent subject: https://github.com/FereshteShakeri/few-shot-MedVLMs .

Keywords: Medical VLMs · Few-shot Learning · Efficient Adaptation

1 Introduction

Deep neural networks have attracted paramount attention in the last decade in the medical image analysis community [20]. Their breakthrough developments in natural image recognition tasks have been successfully applied to a breadth of medical tasks, such as radiology image classification [12], tumor grading in gigapixel stained histology images [28], or diabetic retinopathy grading [5], among others. However, the limitations of such models have restricted their widespread adoption in real clinical settings. In particular, they require large labeled datasets for training reliable task-specific solutions, a burden for medical domains [4], in

* Equal contributions. Corresponding author: fereshteh.shakeri.1@etsmtl.net

which annotated data is usually scarce. In addition, the large domain drifts existing in medical image analysis from inter-scanner, inter-stain, or inter-population variability require continuous adaptation, ideally done in a data-efficient way, *i.e.* using small numbers of labeled samples, a.k.a *few-shot* adaptation. A potential alternative for such adaptation is transfer learning of large pre-trained models that extract robust features. Although popular in computer vision, transferring such models from natural to medical images did not achieve the expected gains [24], due to the fine-grained nature of medical images.

A paradigm shift in transfer learning is currently underway, focused on large-scale pre-training on heterogeneous datasets, which have shown improved transferability, the so-called *foundation models*. In particular, vision-language models, such as CLIP [23] and ALIGN [13], exhibit remarkable adaptability to various downstream tasks. These models can integrate large-scale sources with text supervision (*e.g.* 400M image-text pairs for CLIP), and train joint embedding representations of such modalities by contrastive learning, which have shown astonishing robustness to domain drifts [23]. In addition, such pre-trained knowledge can be efficiently transferred to downstream tasks, in low-shot regimes. Although those are conditions largely desired in the medical-imaging community [21], the direct application of CLIP models has been limited in this domain, since they lack fine-grained expert's medical knowledge.

To alleviate this issue, a myriad of recent works have gathered large open-access medical datasets to build specialized medical vision-language models for radiology [31, 32, 35], histology [10, 11], or ophthalmology [27]. With the current endeavors towards developing and adapting such models to downstream tasks, nevertheless, there are important specific challenges inherent to clinical domains, which are largely being overlooked. First, current studies on medical VLMs predominantly revolve around fine-tuning models with a reduced percentage of the available datasets (e.g., 1% or 10% in [11] or [35]), which still amount to hundreds or thousands of annotated samples. This assumes large labeled datasets for adaptation, which might be inconvenient in clinical applications, particularly when dealing with rare, low-prevalence diseases. Second, pre-training medical foundation models will potentially involve the use of private sources of clinical records, both images and text reports. While recent studies have warned about the potential leaking of the source data from solely using the pre-trained weights [29], fine-tuning the entire encoders during adaptation is still a dominant choice in the literature [32]. Moreover, foundation models tend to improve performances by increasing substantially the number of trainable parameters, thereby requiring substantial hardware requirements for full fine-tuning, which may be unpractical in clinical institutions, with limited computational sources.

Linear probing (LP) is a standard adaptation method, which was also evaluated in the seminal CLIP paper [23]. It is a computationally efficient fine-tuning baseline, which operates in black-box settings, *i.e.* it does not require access to the inner representations of the pre-training models. It consists of updating the weights of a linear classifier on top of the frozen vision encoder, by optimizing the cross-entropy loss built with a few labeled samples in the target task. Un-

fortunately, LP has often been reported as a very weak baseline in the recent literature on few-shot VLMs [23,34,37], as it completely omits the text encoder's knowledge, potentially over-fitting the few labeled images. This has motivated intensive recent research efforts in computer vision, targeted at building convoluted *prompt learning* [3,33,36,37] or *feature-adaptation* [7,34] strategies, which account for such information. In particular, prompt learning is gaining wide popularity in the field. This parameter-efficient family of methods improves adaptation by optimizing the best text input for a target task, via learnable continuous prompts. We demonstrate that such prompt-tuning approaches offer limited performance gains in few-shot medical-image classification, at the cost of imposing an overlooked extensive computational and memory overhead, requiring gradient back-propagation throughout the entire text encoder. Moreover, the assumption of accessing the learned parameters of the text encoder may hinder their deployment in low-resource and privacy-preserving black-box scenarios, which are crucial considerations in medical domains. To address these issues, a few, very recent studies in computer vision have incorporated knowledge from the text encoder to enhance the linear-probe baseline [9, 19].

Given the continuous emergence of foundation models in medical imaging, along with the potential deployment of these popular adaptation methods, we aim at paving the way towards more realistic adaptation of medical VLMs, taking into account transferability scenarios with access to limited labeled examples per task, *i.e.* up to 16 shots. Our main contributions could be summarized as follows:

- We introduce the first structured benchmark for adapting medical vision-language models (VLMs) in a strict few-shot regime.
- We evaluate a simple generalization of the LP baseline, which seeks an optimal blending of the visual prototypes and text embeddings via learnable class-wise multipliers. Surprisingly, such a text-informed LP yields competitive performances in comparison to convoluted prompt-learning and adapter-based strategies, while running considerably faster and accommodating the black-box setting (as it requires access to the output embeddings only).
- We report extensive evaluations and comparisons over three different medical modalities and specialized foundation models, nine downstream tasks and several state-of-the-art few-shot adaptation strategies.

2 Related Work

Prompt learning. One of the foremost categories of approaches in the few-shot adaptation of vision-language models is prompt learning, motivated by the observation that the choice of input prompt may affect the performance of zero-shot prediction. Following the burgeoning interest in prompt learning within the NLP community [8, 14, 26], notable work by [37] introduced context optimization (CoOp) for vision-language models. In CoOp, text is represented as learnable continuous vectors, which are trained as task-specific prompts through few-shot training examples and a standard supervised classification loss. The innovative idea of CoOp has spurred an extensive body of literature on prompt

learning for few-shot vision-language models, yielding numerous sophisticated extensions [3,33,36,38]. For example, CoCoOp [36] additionally learns instance-conditional contexts conditioned on the inputs to improve the generalization of CoOp to unseen classes. PLOT [3] learns multiple prompts to describe each class's characteristics through the minimization of an optimal-transport distance. KgCoOp [33] enhances CoOp's performance on unseen classes by minimizing the discrepancy between the text embeddings generated by the learned prompts and hand-crafted ones. ProGrad [38] aligns few-shot downstream knowledge with large-scale general knowledge, thereby mitigating overfitting the few-shot samples. Given the popularity of prompt-learning methods in vision and NLP, there is currently an emergent interest in their application within the medical field. This includes, for instance, parameter-efficient medical image segmentation [6] and prompt learning on large clinical language models [30], the latter being more closely related to our setting.

Black-box Adapters. Adapters represent another category of approaches within the realm of few-shot adaptation for VLMs. These methods focus on non-linear transformations applied to the pre-trained vision and text features [7,34]. They are multi-layer modules added to the encoder's bottleneck, and whose parameters are fine-tuned over a few-shot task by optimizing the cross-entropy loss. For instance, CLIP-Adapter [7] incorporated a multi-layer perceptron to learn new features, which are blended with the original pre-trained features through residual connections. Tip-Adapter [34] integrated a non-linear, quadratic-complexity module to assess pairwise similarities between the features of the labeled samples, and blended the resulting class scores with the textual features. This category of approaches effectively alleviates the limitation of prompt-learning methods in terms of computational complexity, by eliminating the need for back-propagation over the text encoder. However, their performance relies heavily on key hyper-parameters, particularly those governing the blending between vision and textual features, which require computationally intensive grid searches.

3 Methods

The few-shot image classification setting. Following on from the popular few-shot setting in computer vision [34,37], our approach involves a foundation model pre-trained on a large dataset composed of image-text pairs. The objective is to predict the labels of samples from previously unseen target datasets, via fine-tuning on a limited number of labeled samples, a.k.a *the support set*. For each support image $\boldsymbol{x}_i$, one may compute its vision embedding $\boldsymbol{f}_i = \boldsymbol{\theta}_v(\boldsymbol{x}_i)$, with $\boldsymbol{\theta}_v$ denoting the frozen pre-trained visual encoder. Also, for each given target class $k \in 1, \ldots, K$, one may use a textual description of the class (or a prompt), $\boldsymbol{z}_k$, e.g., "an image of a [class$_k$]", where [class$_k$] is the class name/description. Let $\boldsymbol{t}_k = \boldsymbol{\theta}_t(\boldsymbol{z}_k)$ denotes the corresponding text embedding, and $\boldsymbol{\theta}_t$ the text encoder.

The standard linear-probe (LP) baseline. The standard linear probe (LP), initially evaluated as a few-shot adaptation baseline in the CLIP paper [23], is a linear classifier that exclusively utilizes the frozen vision features. It optimizes

the following cross-entropy loss w.r.t. the last-layer weights of the vision encoder (i.e., the class prototypes), $\mathbf{w} = (\boldsymbol{w}_k)_{1\leq k\leq K}$:

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

where y_{ik} denotes one-hot encoded label of support image $\boldsymbol{x}_i$, i.e., $y_{ik} = 1$ if $\boldsymbol{x}_i$ belongs to class k and 0 otherwise. Unlike prompt learning methods and Adapters, which integrate text knowledge, a limitation of this standard LP baseline is that it omits completely information from the text encoder, *i.e.* $\mathbf{t} = (\boldsymbol{t}_k)_{1\leq k\leq K}$, yielding significantly lower performances than zero-shot predictions [23].

Text-driven linear probe (LP+text). We evaluate a simple generalization of the LP baseline, which we introduced recently in the context of natural-image few-shot tasks [9]. Our method integrates text knowledge while accommodating the black-box setting. It seeks an optimal blending of the visual prototypes and text embeddings via learnable class-wise multipliers, $\boldsymbol{\alpha} = (\alpha_k)_{1\leq k\leq K}$, by optimizing the following loss function:

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

During few-shot adaptation, visual class prototypes $\mathbf{w} = (\boldsymbol{w}_k)_{1\leq k\leq K}$ and class-wise blending parameters $\boldsymbol{\alpha} = (\alpha_k)_{1\leq k\leq K}$ are updated via full-batch gradient descent, while text embeddings $\mathbf{t} = (\boldsymbol{t}_k)_{1\leq k\leq K}$ are kept fixed. To minimize (2), we follow the computationally efficient, full-batch optimizer in [9], in which step sizes are implicit (derived from the Lipschitz-gradient properties of the objective function [9]). This relaxes intensive validation searches for the optimization hyper-parameters, unlike standard gradient descent practices where learning rates are intensively searched over validation sets. Therefore, it runs significantly faster than state-of-the-art few-shot adaptation methods for VLMs.

4 Experiments

Medical vision-language models (VLMs). A comprehensive assessment of the potential of medical VLM adaptation is carried out across three different popular medical domains: histology, radiology, and ophthalmology. In each domain, we utilize an open-access specialized foundation VLM. **Histology**: we employed Quilt-1M [11], with ViT-B/32 vision and GPT2 text encoder. **Ophtalmology**: we utilized FLAIR [27], a foundation model focused on color fundus image understanding. **Radiology**: we focused on chest X-ray (CXR) scans, which have attracted the attention of a large body of literature [31, 32, 35]. Concretely, we used MedCLIP [31] pre-trained on CheXpert [12] and MIMIC-CXR [16] datasets. Since these datasets are also further used for evaluation, we pre-trained this model to control test partition better and avoid test-data

leakage. We followed [31] implementation details. Note that FLAIR and MedCLIP present a similar dual-encoder architecture: ResNet-50 as vision encoder, and BioClinicalBERT [2] text encoder. It is worth mentioning that those models cover a wide range of architectures, both convolutional and ViTs.

Adaptation tasks. Our benchmark encompasses a wide number of downstream tasks for the adaptation of medical VLMs. To ensure a logical transfer of the pre-trained features, each specialized foundation model is used uniquely for datasets from their respective domain. In addition, such open-access datasets are carefully selected to avoid test data leaking, *i.e.* evaluating with data used for pre-training. **Histology**: involve three different organs and cancer types. Concretely, colorectal adenocarcinoma samples in NCT-CRC [17], prostate cancer grading in SICAPv2 [28], and SkinCancer [18]. **Ophtalmology**: we consider MESSIDOR [5] focused on diabetic retinopathy (DR) grading, and FIVES [15] and ODIR200x3 [1], for inter-diseases discrimination. **Radiology**: following the same evaluation benchmark as in [31], we employed $\text{CheXpert}_{5\times200}$ [12], $\text{MIMIC}_{5\times200}$ [16], and RSNA [25]. These datasets include a heterogeneous variety of fine-grained findings, such as pneumonia, atelectasis, edema, or pleural effusion.

Few-shot adaptation protocol and evaluation. Transfer learning from the large-scale pre-trained models is performed in a challenging, but realistic medical setting, in which only a few samples, *i.e.* shots, are available. Following relevant literature in natural image [7, 23, 37], the training subset consists of $S = \{1, 2, 4, 8, 16\}$ images per class randomly sampled for each dataset in all scenarios. To guarantee fair comparisons among different approaches, we deploy a few-shot validation set with the same number of samples for hyper-parameters tuning. We employed the test splits from the original datasets, if available, or performed a 20% hold-out partition otherwise. The evaluation metric is a balanced average accuracy (ACA), widely employed in CXR [31] and Ophthalmology [27] benchmarks. The evaluation is carried out through 5 random seeds to account for the variability in the few shots selected.

Implementation details and baselines. We conduct a comprehensive comparison of several state-of-the-art methods in the few-shot efficient transfer learning of CLIP-based models. Our benchmarks include Zero-shot prediction (*i.e.* no adaptation), Prompt Learning, and black-box Adapter methods. **Zero-shot**: following CLIP [23], these predictions are obtained by computing the softmax cosine similarity between image and text embeddings. Text embeddings for each category are obtained following the specific prompts used in each original VLM's publication. This is, prompt ensembles for MedCLIP [31], and domain-expert descriptions for FLAIR [27] and Quilt-1M [11]. It is worth mentioning that the same text-driven prompts are used when required in other Adapters. **Prompt Learning**: we resort to the popular CoOp [37] and KgCoOp [33]. **Black-box Adapters**: The firstly proposed linear probing in CLIP paper, LP, is considered as a baseline. Concretely, logistic regression is trained with the L-BFGS [22] optimizer. Also, more recent adaptation techniques such as CLIP-adapter [7] and TIP-adapter [34] are included. For TIP-Adapter, we employed its fine-tuned version, TIP-Adapter-F, and set α and β to 1 initially. Later, we find best values

of α and β using the validation set. Finally, we include the efficient proposed LP+text in the benchmark, following its description in Section 3.

Results. Figure 1 shows a quantitative comparison of all studied few-shot adaptation methods on the 9 benchmarks. As demonstrated by the figure LP+text performs relatively well in most cases, outperforming prompt learning methods by a large margin and performing on par with Adapters. In Table 1 we present specific numerical results for each method, averaged per modality. Specific numeric results per dataset are provided in Supp. Materials. It is worth mentioning from the results that Prompt Learning methods rarely outperform black-box Adapters. For instance, the most recent method of such a family, KgCoOp [33] ranges performance drops (*e.g.* [1.3, 3.4]% for S=16) compared with the proposed LP+text. In addition, the significant standard derivation of prompt learning is relatively large, especially in low-shot settings, which motivates the use of Adapters as a more appealing alternative. Comparing the proposed LP+text with other Adapters, our method shows consistent performance gains to the popular TIP-Adapter [34], and performs at par with CLIP-Adapter, albeit being much more computationally efficient, as we later discuss. Finally, while the basic LP suffers a consistent performance drop in the extreme-low data regime (*i.e.* S=1), introducing text information in LP+text prevents it.

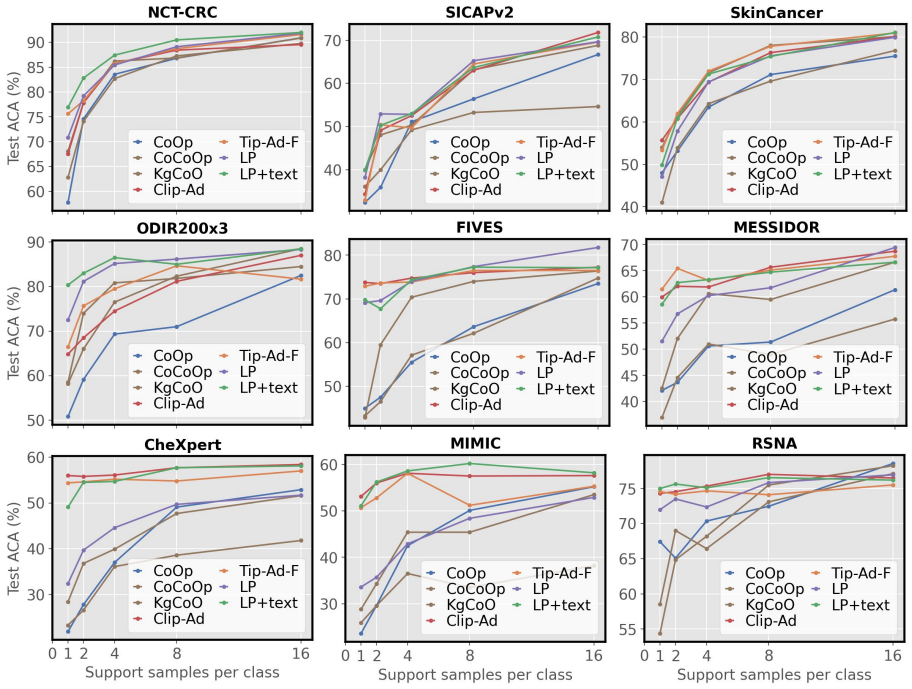

Fig. 1: Comparison of different adaptation methods of Medical VLMs evaluated on 9 benchmarks, averaged over 5 tasks.

Table 1: **Comparison of state-of-the-art methods.** Average ACA (%) on 3 benchmarks for each modality. Best values are highlighted in **bold**.

(a) Histology	S=**1**	S=**2**	S=**4**	S=**8**	S=**16**
Zero-shot ICML'21 [23]			48.33		
CoOp IJCV'22 [37]	46.05 ± 9.79	54.55 ± 8.53	66.04 ± 5.15	71.45 ± 5.53	77.69 ± 1.32
CoCoOp CVPR'22 [3]	46.63 ± 7.71	55.98 ± 5.65	65.39 ± 3.04	70.04 ± 3.05	73.73 ± 2.83
KgCoOp CVPR'23 [33]	53.96 ± 5.95	62.44 ± 3.49	69.37 ± 3.33	76.01 ± 2.46	79.91 ± 1.07
CLIP-Adapter IJCV'23 [7]	52.50 ± 8.31	62.58 ± 4.07	69.21 ± 4.44	75.92 ± 2.48	80.47 ± 1.31
Tip-Adapter-F ECCV'22 [34]	53.97 ± 6.11	63.54 ± 3.41	69.11 ± 4.24	**77.01** ± 2.52	80.69 ± 1.42
Linear probe (LP)	52.05 ± 4.66	63.33 ± 3.24	69.22 ± 4.02	76.64 ± 1.66	80.47 ± 1.61
LP+text [9]	**55.60** ± 6.26	**64.69** ± 3.65	**70.56** ± 3.94	76.52 ± 2.44	**81.26** ± 1.76

(b) Ophtalmology	S=**1**	S=**2**	S=**4**	S=**8**	S=**16**
Zero-shot ICML'21 [23]			65.74		
CoOp IJCV'22 [37]	45.98 ± 12.26	50.11 ± 12.29	58.48 ± 11.12	62.00 ± 6.96	72.45 ± 2.04
CoCoOp CVPR'22 [3]	47.87 ± 12.07	59.19 ± 7.97	69.16 ± 5.79	71.94 ± 4.43	77.16 ± 3.01
KgCoOp CVPR'23 [33]	46.23 ± 10.26	55.03 ± 8.01	62.98 ± 4.49	64.31 ± 4.92	71.67 ± 4.98
CLIP-Adapter IJCV'23 [7]	66.18 ± 4.54	68.00 ± 4.29	70.38 ± 5.90	74.27 ± 3.99	77.65 ± 2.72
Tip-Adapter-F ECCV'22 [34]	66.95 ± 4.03	**71.57** ± 3.78	72.16 ± 3.92	75.42 ± 4.12	75.30 ± 3.38
Linear probe (LP)	64.39 ± 5.57	69.18 ± 5.28	73.13 ± 4.38	75.09 ± 4.24	**79.83** ± 2.34
LP+text [9]	**69.56** ± 6.22	71.15 ± 4.95	**74.72** ± 3.80	**75.66** ± 3.42	77.42 ± 2.07

(c) Radiology	S=**1**	S=**2**	S=**4**	S=**8**	S=**16**
Zero-shot ICML'21 [23]			60.37		
CoOp IJCV'22 [37]	37.64 ± 6.82	40.82 ± 6.76	49.95 ± 6.15	57.21 ± 3.97	62.21 ± 4.00
CoCoOp CVPR'22 [3]	34.52 ± 6.50	40.35 ± 5.63	46.93 ± 6.60	49.19 ± 4.55	52.73 ± 3.46
KgCoOp CVPR'23 [33]	38.57 ± 7.47	46.70 ± 7.11	50.57 ± 5.72	55.39 ± 3.47	60.73 ± 3.51
CLIP-Adapter IJCV'23 [7]	**61.13** ± 2.43	**62.10** ± 2.66	**63.17** ± 2.93	64.06 ± 2.48	**64.15** ± 2.27
Tip-Adapter-F ECCV'22 [34]	59.88 ± 2.80	60.52 ± 1.68	62.64 ± 4.55	60.03 ± 3.29	62.59 ± 2.47
Linear probe (LP)	45.98 ± 4.87	49.63 ± 4.50	53.28 ± 4.80	57.97 ± 3.12	60.50 ± 4.76
LP+text [9]	58.39 ± 5.03	**62.10** ± 3.80	62.79 ± 3.19	**64.80** ± 2.79	**64.15** ± 3.20

Assessing computational workload. Here we evaluate the efficiency of the methods considered by presenting their computational overhead. We also indicate whether these methods enable black-box adaptation, which is a crucial consideration for addressing practical, real-world demands. Furthermore, we outline the number of parameters to be learned during training as an indicator of model complexity. This comparison Table 2 shows that, beyond outperforming state-of-the-art methods as shown in previous sections, LP+text stands out as the most efficient method. Complementary, it is worth noting that LP+text uses around 800MB of peak GPU memory, whereas CoCoOP requires up to 28GB (based on NCT-CRC experiments). This makes prompt learning methods inefficient for institutions with limited access to high-resource GPUs.

Conclusions. Inspired by the computer vision field, we established a new few-shot adaption setting for medical VLMs. We also introduced a generalization of LP baseline, integrating image and text embeddings through learnable class-wise multipliers. Evaluations across various benchmarks show that the proposed

Table 2: **Computational Efficiency.** Experiments on a single 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 [23]		n/a	✓	n/a
CoOp [37]	*Prompt-Learning*	3min	✗	$K \times n_{ctx1} \times D$
CoCoOp [36]		12min	✗	$n_{ctx2} \times D + C$
KgCoOp [33]		3min	✗	$K \times n_{ctx1} \times D$
Clip-Adapter [7]	*CLIP-based Adapters*	2min	✓	$2(D_1 \times D_2)$
Tip-adapter-F [34]		2min	✓	$K \times S \times D$
LP	*Linear probe*	43s	✓	$K \times D$
LP+text [9]		4s	✓	$K(D+1)$

method stands out for its performance in different scenarios, its simplicity, computational efficiency, and its potential applicability in black-box scenarios.

Acknowledgments. This work was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and Montreal University Hospital Research Center (CRCHUM). We also thank Calcul Quebec and Compute Canada.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article

References

1. Ocular disease intelligent recognition (odir) (2019), https://odir2019.grand-challenge.org/
2. Alsentzer, E., et al.: Publicly available clinical BERT embeddings. In: Clinical Natural Language Processing Workshop (2019)
3. Chen, G., et al.: Prompt learning with optimal transport for vision-language models. In: International Conference on Learning Representations (2023)
4. Chen, X., et al.: Recent advances and clinical applications of deep learning in medical image analysis. Medical Image Analysis **79** (2022)
5. Decencière, E., et al.: Feedback on a publicly distributed image database: The messidor database. Image Analysis & Stereology **33**, 231–234 (07 2014)
6. Fischer, M., Bartler, A., Yang, B.: Prompt tuning for parameter-efficient medical image segmentation. Medical Image Analysis **91**, 103024 (2024)
7. Gao, P., et al.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision **132**, 581–595 (2023)
8. hong, Z., Friedman, D., Chen, D.: Factual probing is [mask]: Learning vs. learning to recall. In: Conference of the North American Chapter of the Association for Computational Linguistics (2021)
9. Huang, Y., Shakeri, F., Dolz, J., Boudiaf, M., Bahig, H., Ayed, I.B.: Lp++: A surprisingly strong linear probe for few-shot clip. In: IEEE Conference on Computer Vision and Pattern Recognition (2024)
10. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature Medicine **29**, 1–10 (2023)

11. Ikezogwo, W.O., et al.: Quilt-1m: One million image-text pairs for histopathology. In: Neural Information Processing Systems (2023)
12. Irvin, J., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI (2019)
13. Jia, C., et al.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning (2021)
14. Jiang, Z., Xu, F., Araki, J., Neubig, G.: How can we know what language models know. In: Association for Computational Linguistics (2020)
15. Jin, K., et al.: Fives: A fundus image dataset for artificial intelligence based vessel segmentation. Scientific Data **9** (2022)
16. Johnson, A.E., et al.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6** (2019)
17. Kather, J.N., Halama, N., Marx, A.: 100,000 histological images of human colorectal cancer and healthy tissue. Zenodo **5281** (2018)
18. Kriegsmann, K., et al.: Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections. Frontiers in Oncology **12** (2022)
19. Lin, Z., Yu, S., Kuang, Z., Pathak, D., Ramanan, D.: Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19325–19337 (2023)
20. Litjens, G., et al.: A survey on deep learning in medical image analysis. Medical Image Analysis **42** (2017)
21. Moor, M., et al.: Foundation models for generalist medical artificial intelligence. Nature **616**, 259–265 (4 2023)
22. Nocedal, J.: Updating quasi-newton matrices with limited storage. Mathematics of Computation **35**(151), 773–782 (1980)
23. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021)
24. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. In: Advances in neural information processing systems (2019)
25. Shih, G., et al.: Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiology: Artificial Intelligence **1** (2019)
26. Shin, T., et al.: Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In: CoRR (2020)
27. Silva-Rodriguez, J., Chakor, H., Kobbi, R., Dolz, J., Ayed, I.B.: A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. ArXiv Preprint (2023)
28. Silva-Rodríguez, J., Colomer, A., Sales, M.A., Molina, R., Naranjo, V.: Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. Computer methods and programs in biomedicine **195** (2020)
29. Song, C., Ristenpart, T., Shmatikov, V.: Machine learning models that remember too much. In: Conference on Computer and Communications Security (2017)
30. Taylor, N., et al.: Clinical prompt learning with frozen language models. IEEE Transactions on Neural Networks and Learning Systems (2023)
31. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Empirical Methods in Natural Language Processing (2022)

32. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In: International Conference on Computer Vision (2023)
33. Yao, H., Zhang, R., Xu, C.: Visual-language prompt tuning with knowledge-guided context optimization. In: IEEE Conference on Computer Vision and Pattern Recognition (2023)
34. Zhang, R., et al.: Tip-adapter: Training-free adaption of clip for few-shot classification. In: European Conference on Computer Vision (2022)
35. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: MHLC (2022)
36. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: IEEE Conference on Computer Vision and Pattern Recognition (2022)
37. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**, 2337–2348 (2022)
38. Zhu, B., Niu, Y., Han, Y., Wu, Y., Zhang, H.: Prompt-aligned gradient for prompt tuning. In: International Conference on Computer Vision (2023)