This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Few-Shot, Now for Real: Medical VLMs Adaptation without Balanced Sets or Validation

Julio Silva-Rodríguez[1,✉], Fereshteh Shakeri[1,2], Houda Bahig[2], Jose Dolz[1,2], and Ismail Ben Ayed[1,2]

[1]ÉTS Montréal
✉julio-jose.silva-rodriguez@etsmtl.ca
[2]Centre de Recherche du Centre Hospitalier de l'Université de Montréal (CRCHUM)

Abstract. Vision-language models (VLMs) are gaining attention in medical image analysis. These are pre-trained on large, heterogeneous data sources, yielding rich and transferable representations. Notably, the combination of modality-specialized VLMs with few-shot adaptation has provided fruitful results, enabling the efficient deployment of high-performing solutions. However, previous works on this topic make strong assumptions about the distribution of adaptation data, which are unrealistic in the medical domain. First, prior art assumes access to a balanced support set, a condition that breaks the natural imbalance in disease prevalence found in real-world scenarios. Second, these works typically assume the presence of an additional validation set to fix critical hyper-parameters, which is highly data-inefficient. This work challenges these favorable deployment scenarios and introduces a realistic, imbalanced, validation-free adaptation setting. Our extensive benchmark across various modalities and downstream tasks demonstrates that current methods systematically compromise their performance when operating under realistic conditions, occasionally even performing worse than zero-shot inference. Also, we introduce a training-free linear probe that adaptively blends visual and textual supervision. Detailed studies demonstrate that the proposed solver is a strong, efficient baseline, enabling robust adaptation in challenging scenarios. Code is available: https://github.com/jusiro/SS-Text .

Keywords: Medical VLMs · Few-shot adaptation · Realistic assessment

1 Introduction

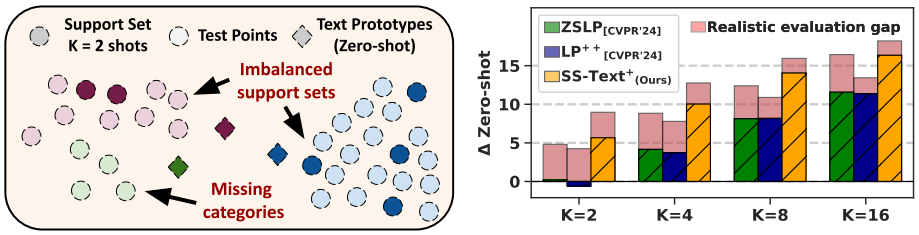
The recent advancements in deep learning have yielded remarkable outcomes to enhance computer-aided medical image analysis [18]. Nevertheless, these have been classically hindered by the necessity of using large, labeled datasets for training highly specialized solutions [3]. Currently, foundation models are driving a paradigm shift, enabling more data-efficient, robust solutions [22,37]. Particularly, specialized contrastive vision-language models (VLMs) are being developed in the primary medical image modalities, i.e., radiology [38,34], histology [10,19] or retina [30]. These models are pre-trained on large datasets through text supervision, which enables the leveraging of heterogeneous data sources that encompass different tasks and domains. Thanks to such pre-training, they exhibit

remarkable adaptability to various downstream tasks. For example, VLMs enable zero-shot predictions by leveraging text knowledge into the learned joint multi-modal space and have shown astonishing robustness to domain drifts [25]. However, recent studies have underscored that such zero-shot performance highly depends on the target concept's frequency [32], and domain [21] during pre-training. This motivates the integration of supervisory signals for the target tasks, ideally in a data-efficient manner, i.e., using small numbers of labeled samples, also known as few-shot adaptation.

Few-shot adaptation is of special interest in medical applications due to the scarcity of data resources or when tackling low-prevalence diseases. From a practical standpoint, a practitioner would gather a few examples for each category, e.g., 4, 8, or 16, and efficiently adapt a pre-trained model. Indeed, VLMs are strong few-shot learners [25], particularly in the efficient black-box scenario [23], i.e., transferring the embedded representations. A recent body of literature has focused on improving this transfer using the so-called Adapters [6,36,35,17,29,9], which combine visual and textual supervision. Consequently, due to its success in natural image domains, recent work in [27] has deployed this setting in modern modality-specialized medical VLMs. However, we argue that the scenarios explored in natural domains are hardly directly applicable to medical image analysis, where data access presents more challenging constraints.

First, these works assume access to a balanced support set for adaptation, such that the number of shots is homogeneous across categories. However, this strategy is unrealistic in medical domains, where disease prevalence is naturally imbalanced. For example, in tumor grading tasks [28] or diabetic retinopathy grading [5], gathering examples for the severe disease stages is typically challenging. Indeed, in low-data regimes, there may be potentially missing categories in the support set (the least prevalent ones). Thus, evaluating such few-shot Adapters using balanced support sets provides an overoptimistic, biased assessment of the potential of these techniques in real-world scenarios. Note that this is not an anecdotal problem, but its effects are being observed in recent literature. For example, the authors of Quilt [11] reported better performance in low-data but balanced regimes than using the whole dataset for adaptation when transferring histology VLMs. Moreover, if evaluated in realistic scenarios (see Fig. 1(b)), current SoTA methods suffer significant performance degradation.

Second, current few-shot adaptation strategies suffer from data-inefficient model selection strategies, i.e., how critical hyper-parameters and learning schedules are fixed. Early works even leveraged significant data sources for this task [6,36], a limitation smoothed by ulterior literature [17,36]. Still, current medical benchmarks assume a few-shot validation set [27], which doubles the required samples for adaptation. Again, such a scenario poses an additional burden in medical applications. Therefore, we argue that few-shot Adapters should rely uniquely on a single support set and incorporate appropriate model selection strategies. Indeed, as we later demonstrate in Fig. 3(b), using such additional validation data for training brings larger performance gains.



(a) Realistic few-shot sets **(b) Performance drop**

Fig. 1: **Realistic adaptation of medical VLMs**: disease prevalence might produce imbalanced support sets or missing categories during low-shot adaptation (a), producing severe performance detriments in SoTA black-box Adapters (b).

The main contributions of this paper can be summarized as:

- We highlight the limitations of current few-shot Adapters for medical VLMs, which often degrade its performance (Fig. 1(b)) in realistic scenarios that account for real-world disease prevalence and constraints to data access.
- Hence, we introduce two validation-free scenarios for assessing its few-shot adaptation with imbalanced support sets and potential missing categories in the low-data regime (Fig. 1(a)).
- We propose SS-Text$^+$, a training-free, constraint linear probe that adaptively leverages imbalanced visual and textual information for each category, providing more robust performance than the current state-of-the-art (SoTA).

2 Related Work

Transfer learning in VLMs. Initial efforts were devoted to Prompt Learning [39,7], which learns a set of input tokens to optimize the best textual prototypes. However, current literature has underscored its inefficiency compared to black-box Adapters [6,36,9,29], which directly operate over pre-computed features. While their performance is comparable, they are orders of magnitude faster [9,27]. We follow up on [27] and focus on assessing these efficient Adapters.

Black-box Adapters. Current methods are linear probes that combine textual and visual information. Indeed, a simple linear probe with proper textual-based initialization is already a strong baseline adaptation [29]. More complex methods can be categorized into ensemble-based [6,36,9,27], which combine visual and textual similarities into the final output, and constraint-based [17,29], which enforce the learned probes to remain close to the initial textual class embeddings. For example, the CrossModal linear probe [17] employs textual embeddings as additional support samples for logistic regression, while CLAP [29] applies explicit penalties. Our proposed solver belongs to the second family of methods and provides a training-free, flexible solver designed for imbalanced scenarios.

3 Methods

3.1 Contrastive vision-language models

Zero-shot. Contrastive VLMs project data points into an ℓ_2-normalized D-dimensional shared embedding space, yielding the corresponding visual, $\mathbf{v} \in \mathbb{R}^{D\times 1}$, and text, $\mathbf{t} \in \mathbb{R}^{D\times 1}$, embeddings. Given a pre-computed image feature and class-wise prototypes, $\mathbf{W} = (\mathbf{w}_c)_{1\leq c\leq C}$, with $\mathbf{w}_c \in \mathbb{R}^{D\times 1}$, and C the number of target categories, probabilities can be computed as:

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

where τ is a pre-trained temperature parameter, $\mathbf{v}^\top \mathbf{w}$ is the cosine similarity, and $\mathbf{p}(\mathbf{W}) = (p_c(\mathbf{W}))_{1\leq c\leq C}$ corresponds to the predicted probabilities vector. Contrastive VLMs allow zero-shot predictions, i.e., no need to learn $\mathbf{W}$, by embedding a textual description for each label. Concretely, given a set of J textual embeddings for each target category, $\{\{\mathbf{t}_{cj}\}_{j=1}^{J}\}_{c=1}^{C}$, the zero-shot prototypes are the average of the text embeddings for each class, $\mathbf{t}_c = \frac{1}{J}\sum_{j=1}^{J} \mathbf{t}_{cj}$. These class weights are then applied to Eq. (1), i.e., $\mathbf{W}^* = (\mathbf{t}_c)_{1\leq c\leq C}$.

3.2 Towards realistic few-shot adaptation scenarios

Let us define an adaptation set, $\mathcal{D}_{\text{adapt}} = \{(\mathbf{v}_n, \mathbf{y}_n)\}_{n=1}^{N}$, composed of labeled embedded visual examples. Note that $\mathbf{y}$ is the one-hot encoding of the label space, $\mathcal{Y} = \{1, 2, ..., C\}$. Typically, in a few-shot support set, N takes small values, defined by the number of examples available for each category, e.g., $K \in \{1, 2, 4, 8, 16\}$, such that the total number of samples is $N = K \times C$. In the following, we detail the standard adaptation scenario explored in prior art, and two more challenging scenarios we propose for addressing real-world constraints:

- ***i*) Standard**. The number of shots for each category, $\mathbf{K} = (K_c)_{1\leq c\leq C}$, is constant, such that $K_c = N/C$. However, this assumption does not account for the realistic imbalance that naturally arises from variable disease prevalence in specific demographics or institutions.
- ***ii*) Realistic**. Let us define a label-marginal distribution among categories, $\mathbf{m} = (m_c)_{1\leq c\leq C}$, s.t. $\sum \mathbf{m} = 1$. For a specific K, the support set comprises $N = K \times C$ samples, each from a category sampled from the label-marginal distribution, i.e., $y \sim \mathbf{m}$. In this scenario, the number of samples from a given category might be variable and even zero if m_c and K are small, which is a challenging scenario with potentially missing categories.
- ***iii*) Relaxed**. This intermediate scenario assumes that at least one labeled example can be retrieved for each category of interest, i.e. $K_c \geq 1\ \forall c \in C$.

3.3 SS-Text: training-free, text-informed linear probe

Multi-modal black-box Adapters involve linear probes that combine visual and textual information [17,29]. These can be modeled from a constrained optimization perspective, where the weights matrix, $\mathbf{W}$ in Eq. (1), is adjusted minimizing cross-entropy loss, but encouraged to remain near the textual class prototypes:

$$\min_{\mathbf{W}} \ \mathcal{L}(\mathbf{W}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\ \ln(p_{ic}(\mathbf{W})) + \frac{\lambda}{2}\sum_{c=1}^{C}||\mathbf{w}_c - \mathbf{t}_c||_2^2. \tag{2}$$

This objective is typically optimized via gradient descent [29]. However, due to the scarcity of adaptation data, its global minimum might be suboptimal when generalizing to test data. Current literature controls such overfitting by learning procedures based on gradient-based optimizers, which necessitate a fine-grained hyper-parameter search, e.g., for iterations, and learning rate schedule, thereby negating the validation-free scenario explored in this work. To address such limitation, we leverage our previously introduced training-free solver, SS-Text, which was recently presented in [31] for its speed in adaptation. The objective in Eq. (2) is evaluated as the sum of two terms, $\mathcal{L} = g_1 + g_2$, such that:

$$g_1 = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic}\ (\mathbf{v}^\top \mathbf{w}_c/\tau) + \frac{\lambda}{2}\sum_{c=1}^{C}||\mathbf{w}_c - \mathbf{t}_c||_2^2, \tag{3}$$

$$g_2 = \frac{1}{N}\sum_{i=1}^{N}\ln\left(\sum_{j=1}^{C}\exp(\mathbf{v}^\top \mathbf{w}_j/\tau)\right). \tag{4}$$

The solution of Eq. (2) can be approximated by solely addressing the first term, g_1, which is related to a hard-label assignment of the support embeddings. Independent of the value of $\lambda > 0$, g_1 is convex w.r.t. each $\mathbf{w}_c$, as it is the sum of linear and convex functions. Particularly, its minimum is:

$$\mathbf{w}_c^{g_1} = \ \arg\min_{\mathbf{w}_c}\ \frac{\partial g_1}{\partial \mathbf{w}_c} = \frac{1}{\lambda N \tau}\sum_{i=1}^{N} y_{ic}\mathbf{v} + \mathbf{t}_c. \tag{5}$$

This solution is a linear combination of scaled visual and textual embeddings for each class. Its closed-form makes it convenient for validation-free deployments. However, it may not yet handle the imbalanced, realistic scenarios addressed in this work, as visual and textual information are weighted equally across all categories. To overcome such a limitation, we propose **SS-Text**$^+$, a modified version of the solution in [31] for this specific challenge.

First, the general multiplier lambda is fixed adaptively class-wise, such that $\lambda_c = 1/(K_c\tau)$. Thus, the effect of text knowledge decreases as more support data becomes available for a specific category. Note that this choice accommodates the solution to the realistic, imbalanced scenario with missing categories. For example, if no supervision is provided for one particular category ($K_c = 0$), the obtained prototype will still rely on the available textual information.

Second, to avoid prototypes overlapping due to insufficient supervision, e.g., when missing categories and relying primarily on text supervision, we propose a post-processing stage that refines the overall solution. Concretely, the resultant prototypes are distanced one each other such that $\mathbf{w}_c^* = \mathbf{w}_c^{g_1} - \frac{\mathbf{g}_c}{||\mathbf{g}_c||}$, where $\mathbf{g}_c = \mathbf{w}_c^{g_1} - \frac{1}{C-1}\sum_{c' \in \mathcal{Y}'_c} \mathbf{w}_{c'}^{g_1}$, such that $\mathcal{Y}'_c = \{c' \in \mathcal{Y} \mid c' \neq c\}$. This term can be interpreted as repelling $\mathbf{w}_c^{g_1}$ from the prototypes of the other categories.

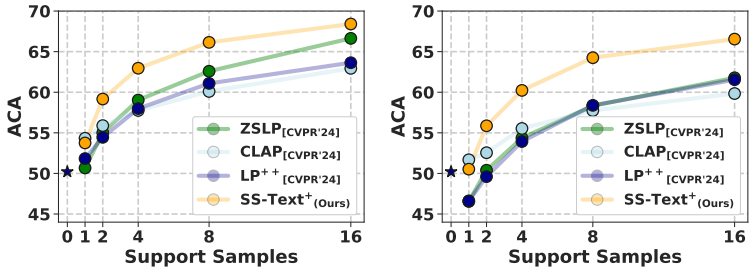
4 Experiments

4.1 Setup

Medical vision-language models. Three modalities are tackled, as in [27]. **Histology**: CONCH [19] is used, which deploys a customized larger-scale ViT-B/16 visual backbone. **Ophtalmology:** FLAIR [30], the expert-knowledge guided VLM for retina imaging, is selected. **Chest X-ray (CXR)**: CONVIRT [38] pre-trained on MIMIC [14] is used. FLAIR and CONVIRT are composed of fine-tuned BioClinicalBERT text encoders [1], and the vision encoder is ResNet-50.
Downstream tasks include tissue/disease classification or grading in balanced (b) and imbalanced (u) scenarios. **Histology**: three different organs are included: colon in NCT-CRC [15] (u), prostate Gleason grading in SICAPv2 [28] (u), and SkinCancer [16] (u). **Ophtalmology**: diabetic retinopathy using MESSIDOR [5] (u), myopic maculopathy staging in MMAC [24] (u), and disease detection in FIVES [13] (b), are included. **CXR**: five CheXpert [12] (b) categories, as in [34], 19 categories in NIH-LT [33,8] (u), and pneumonia types [4,26] (b) are assessed.
Adaptation methods. Relevant gradient-based black-box Adapters are leveraged as baselines. These are trained in a validation-free setting, i.e., using fixed learning hyper-parameters: training for 300 steps using full-batch SGD optimizer with a momentum of 0.9 and an initial learning rate of 0.1 with a cosine-scheduler decay. For vanilla linear probing, we follow ZS-LP initialization [29]. Also, the recent text-informed SoTA CLAP [29] and LP++ [9,27] are integrated. For LP++, the learning rate is data-driven, as proposed by the authors. Finally, previous relevant multi-modal Adapters are assessed: CLIP-Adapter [6] ($\alpha_{\text{clipad}} = 0.2$), TIP-Adapter [36]($\alpha_{\text{tip}}$, $\beta_{\text{tip}} = 1.0$) training-free and fine-tuned (f), TaskRes [35]($\alpha_{\text{taskres}} = 1.0$), and CrossModal linear probe [17].
Evaluation protocol. Support sets with $K \in \{1, 2, 4, 8, 16\}$ are sampled from train data as described in Section 3.2. The label-marginal distribution, $\mathbf{m}$, is estimated from the whole train subset by measuring label frequencies. Class-wise balanced accuracy (ACA) is employed as a figure of merit, as recommended in [20]. All results are the average across 20 different random seeds for sampling.

4.2 Results

Results across shots. Results for the most recent SoTA are depicted in Fig. 2 for the standard (a) and realistic (b) scenarios. First, it is worth noting that the proposed SS-Text$^+$ outperforms relevant SoTA in the validation-free scenario.



(a) Standard setting **(b) Realistic scenario**

Fig. 2: **Results across shots**: comparison of SoTA black-box Adapters.

Second, all methods consistently drop the performance when considering realistic support sets. Note that some methods degrade below Zero-shot (no adaptation) in the low-shot scenario. In contrast, our method is more robust to this decay.

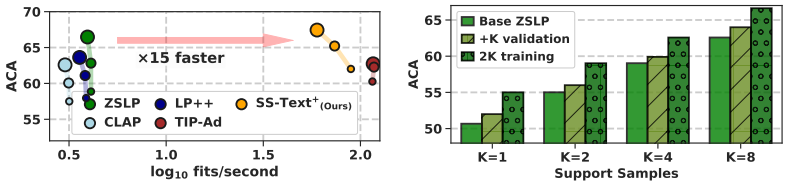
Detailed results per modality. Table 1 introduces the performance for the three explored scenarios per modality. Results are reported with K=4, a low-shot regime showcasing the effects of the missing classes. For example, vanilla linear probing (ZSLP) degrades nearly −3.7 in the relaxed setting and −4.6 in the realistic scenario. This performance drop is alleviated in text-informed solutions, such as LP++, which suffers a drop of −1.9 and −4.1, respectively. Note that the gap is larger in the realistic scenario, where some methods perform below zero-shot for certain modalities, e.g., LP++ in ophthalmology. In contrast, SS-Text$^+$ is more robust and the best solution across all scenarios.

Table 1: **Results per modality and type of support set** with $K = 4$ shots.

	(a) Standard				(b) Relaxed				(c) Realistic			
Method	Hist.	Opht.	CXR	**Avg.**	Hist.	Opht.	CXR	**Avg.**	Hist.	Opht.	CXR	**Avg.**
Zero-shot [25]	61.7	56.8	31.8	50.1	61.7	56.8	31.8	50.1	61.7	56.8	31.8	50.1
CLIP-Adapt [6]	78.1	60.1	42.3	60.2	74.0	58.2	39.5	57.2	73.0	56.7	38.8	56.2
TIP-Adapt [36]	77.4	61.7	43.5	60.8	38.0	37.4	36.8	37.4	38.0	38.2	36.8	37.7
TIP-Adapt(f) [36]	**82.0**	**63.6**	43.7	**63.1**	65.6	52.6	37.7	52.0	65.6	46.8	37.6	50.0
TaskRes [35]	74.7	60.5	41.9	59.0	71.4	56.7	38.0	55.3	71.0	54.4	37.7	54.4
CrossModal [17]	71.0	62.2	44.0	59.1	70.9	60.0	40.1	57.0	70.8	56.3	39.4	55.5
ZSLP [29]	74.7	60.5	41.9	59.0	71.4	56.7	38.0	55.3	71.0	54.4	37.7	54.4
CLAP [29]	66.9	62.6	43.7	57.8	67.5	61.0	40.0	56.2	67.4	59.7	39.5	55.5
LP++ [9]	70.2	61.0	42.7	58.0	68.8	59.3	40.2	56.1	68.6	54.9	38.3	53.9
SS-Text$^+$ (*Ours*)	81.2	63.2	**44.4**	63.0	**76.3**	**62.8**	**44.2**	**61.1**	**75.0**	**61.7**	**44.0**	**60.2**

Efficiency analysis. Fig. 3(a) shows the efficiency of SS-Text$^+$. Note that gradient-based linear probes are already highly efficient due to their black-box nature, especially compared to prompt learning as assessed in [27]. Notably, SS-Text$^+$ is $\geq 15\times$ faster, which showcases its utility in high-latency applications.

Is it worth using a few-shot validation set? Using the vanilla linear probe, Fig. 3(b) provides evidence that the performance gains from using few-shot validation sets are smaller than leveraging this data for training a validation-free approach that follows a fixed setting and does not require model selection.



(a) Efficiency **(b) How to use the validation set?**

Fig. 3: **Efficiency and model selection studies**. Results in the standard scenario. In (a), The dot size indicates the number of shots, i.e., $K \in \{4, 8, 16\}$.

Hyper-parameter analysis. Table 2(a) evaluates the proposed task-adaptive text regularization weight λ_c. Our configuration, which considers the specific amount of supervision received from each category, provides a robust benefit across shots compared to non-adaptive alternatives or performance-driven choices as proposed in CLAP [29]. Also, Table 2(b) shows the benefit of the repelling prototype post-processing, which provides a +0.7 gain in accuracy.

Long-tail training. Popular imbalanced deep learning mechanisms are re-balancing, augmentation, or decoupled learning techniques (see [8]). Due to the black-box nature of few-shot adaptation, the latter ones do not apply to our scenario, and the first, e.g., class weighting (cw), is not applicable if missing categories. Table 2(c) reports the positive effect of re-balancing techniques over vanilla linear probing in the explored relaxed setting. Still, its performance falls behind the proposed text-informed, training-free solution (see Table 2(b)).

Table 2: **Ablation studies** in SS-Text$^+$ and long-tail learning techniques.

Method	$K \rightarrow$	1	2	4	8	16
Text regularization fixed for all tasks.						
$\lambda = 0.1/\tau$		49.1	54.7	59.9	63.7	66.4
$\lambda = 1.0/\tau$		48.2	53.7	57.6	60.6	61.5
$\lambda = 10/\tau$		27.7	29.4	29.3	29.1	29.5
Ours: task-adaptive regularization value.						
λ_c as in [29]		49.6	55.1	60.1	63.4	65.5
$\lambda_c = 1/(K_c\tau)$		**50.5**	**55.9**	**60.2**	**64.2**	**66.5**

(a) λ_c in SS-Text$^+$ (realistic setting)

Method	Hist.	Fundus	CXR	**Avg.**
SS-Text$^+$ w/o repel	**83.2**	67.0	47.2	65.8
SS-Text$^+$	**83.2**	**67.5**	**48.9**	**66.5**

(b) SS-Text$^+$ Components (K=16, realistic)

Method	Hist.	Fundus	CXR	**Avg.**
ZSLP [29]	77.2	63.3	45.1	61.9
w/ cw	79.2	65.4	47.5	64.0
w/ cw+LDAM[2]	**79.6**	**65.6**	**47.7**	**64.3**

(c) Long-tail strategies (K=16, $K_c \geq 1$)

5 Conclusions

Several aspects have been introduced that should be considered when adapting medical vision-language models in the few-shot data regime. These include a more realistic design of the support set and validation-free model selection pipelines. In this context, the proposed training-free SS-Text$^+$ is a strong baseline for measuring progress in the field.

Acknowledgments. This work was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Montréal 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. Alsentzer, E., et al.: Publicly available clinical BERT embeddings. In: Clinical Natural Language Processing Workshop (2019) 6
2. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. In: NeurIPS. vol. 32 (2019) 8
3. Chen, X., Wang, X., Zhang, K., Fung, K.M., Thai, T.C., Moore, K., Mannel, R.S., Liu, H., Zheng, B., Qiu, Y.: Recent advances and clinical applications of deep learning in medical image analysis. Medical Image Analysis **79** (2022) 1
4. Chowdhury, M.E.H., et al.: Can ai help in screening viral and covid-19 pneumonia? IEEE Access **8**, 132665–132676 (2020) 6
5. Decencière, E., et al.: Feedback on a publicly distributed image database: The messidor database. Image Analysis & Stereology **33**, 231–234 (07 2014) 2, 6
6. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision (2023) 2, 3, 6, 7
7. Hantao Yao, Rui Zhang, C.X.: Visual-language prompt tuning with knowledge-guided context optimization. In: CVPR (2023) 3
8. Holste, G., et al.: Long-tailed classification of thorax diseases on chest x-ray: A new benchmark study. In: MICCAI DALI Workshop. pp. 22–32. Springer (2022) 6, 8
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: CVPR. pp. 23773–23782 (2024) 2, 3, 6, 7
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) 1
11. Ikezogwo, W.O., et al.: Quilt-1m: One million image-text pairs for histopathology. In: NeurIPS Datasets and Benchmarks Track (2023) 2
12. Irvin, J., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI. vol. 33, pp. 590–597 (2019) 6
13. Jin, K., et al.: Fives: A fundus image dataset for artificial intelligence based vessel segmentation. Scientific Data **9**, 475 (08 2022) 6

14. Johnson, A., et al.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**, 317 (2019) 6
15. Kather, J.N., Halama, N., Marx, A.: 100,000 histological images of human colorectal cancer and healthy tissue. Zenodo10 **5281** (2018) 6
16. 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**, 1022967 (2022) 6
17. Lin, Z., Yu, S., Kuang, Z., Pathak, D., Ramanan, D.: Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In: CVPR (2023) 2, 3, 5, 6, 7
18. Litjens, G., et al.: A survey on deep learning in medical image analysis. Medical Image Analysis **42** (2017) 1
19. Lu, M.Y., et al.: A visual-language foundation model for computational pathology. Nature Medicine **30**, 863–874 (2024) 1, 6
20. Maier-Hein, L., et al.: Metrics reloaded: recommendations for image analysis validation. Nature Methods **21** (2024) 6
21. Mayilvahanan, P., et al.: In search of forgotten domain generalization. In: ICLR (2025) 2
22. Moor, M., et al.: Foundation models for generalist medical artificial intelligence. Nature **616**, 259–265 (4 2023) 1
23. Ouali, Y., Bulat, A., Martinez, B., Tzimiropoulos, G.: Black box few-shot adaptation for vision-language models. In: ICCV (2023) 2
24. Qian, B., et al.: A competition for the diagnosis of myopic maculopathy by artificial intelligence algorithms. JAMA ophthalmology **142**(11), 1006–1015 (2024) 6
25. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021) 2, 7
26. Rahman, T., et al.: Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. Computers in Biology and Medicine **132**, 104319 (2021) 6
27. Shakeri, F., Huang, Y., Silva-Rodríguez, J., Bahig, H., Tang, A., Dolz, J., Ben Ayed, I.: Few-shot adaptation of medical vision-language models. In: MICCAI. pp. 553–563 (2024) 2, 3, 6, 7
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**, 105637 (2020) 2, 6
29. Silva-Rodríguez, J., Hajimiri, S., Ayed, I.B., Dolz, J.: A closer look at the few-shot adaptation of large vision-language models. In: CVPR. pp. 23681–23690 (2024) 2, 3, 5, 6, 7, 8
30. Silva-Rodríguez, 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. Medical Image Analysis **99**, 103357 (2025) 1, 6
31. Silva-Rodríguez, J., et al.: Full conformal adaptation of medical vision-language models. In: Information Processing in Medical Imaging (IPMI) (2025) 5
32. Udandarao, V., et al.: No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. In: NeurIPS (2024) 2
33. Wang, X., et al.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: CVPR. pp. 3462–3471 (2017) 6
34. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: EMNLP. pp. 1–12 (10 2022) 1, 6

35. Yu, T., Lu, Z., Jin, X., Chen, Z., Wang, X.: Task residual for tuning vision-language models. In: CVPR. pp. 10899–10909 (2023) 2, 6, 7
36. Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. In: ECCV. pp. 1–19 (11 2022) 2, 3, 6, 7
37. Zhang, S., Metaxas, D.: On the challenges and perspectives of foundation models for medical image analysis. Medical Image Analysis **91**, 102996 (2024) 1
38. Zhang, Y., Jiang, H., Miura, Y., Manning, C.D., Langlotz, C.P.: Contrastive learning of medical visual representations from paired images and text. In: Machine Learning for Healthcare (MHLC). pp. 1–24 (2022) 1, 6
39. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision (2022) 3