

Trustworthy Few-Shot Transfer of Medical VLMs through Split Conformal Prediction

Julio Silva-Rodríguez[✉], Ismail Ben Ayed, and Jose Dolz

ÉTS Montréal

✉julio-jose.silva-rodriguez@etsmtl.ca

Abstract. Medical vision-language models (VLMs) have demonstrated unprecedented transfer capabilities and are being increasingly adopted for data-efficient image classification. Despite its growing popularity, its reliability aspect remains largely unexplored. This work explores the split conformal prediction (SCP) framework to provide trustworthiness guarantees when transferring such models based on a small labeled calibration set. Despite its potential, the generalist nature of the VLMs’ pre-training could negatively affect the properties of the predicted conformal sets for specific tasks. While common practice in transfer learning for discriminative purposes involves an adaptation stage, we observe that deploying such a solution for conformal purposes is suboptimal since adapting the model using the available calibration data breaks the rigid exchangeability assumptions for test data in SCP. To address this issue, we propose *transductive split conformal adaptation* (SCA-T), a novel pipeline for transfer learning on conformal scenarios, which performs an unsupervised transductive adaptation jointly on calibration and test data. We present comprehensive experiments utilizing medical VLMs across various image modalities, transfer tasks, and non-conformity scores. Our framework offers consistent gains in efficiency and conditional coverage compared to SCP, maintaining the same empirical guarantees. The code is publicly available: <https://github.com/jusiro/SCA-T>.

Keywords: Medical VLMs · Conformal Prediction · Transduction.

1 Introduction

Vision-language models (VLMs) pre-trained on extensive data sources, such as CLIP [28], are witnessing widespread adoption in a plethora of computer vision problems [34,32], showcasing unprecedented transfer capabilities to downstream tasks. These models have rapidly gained traction in safety-critical scenarios, such as healthcare, where specialized medical VLMs have been recently introduced for radiology [45,42], histology [11,20], or retina [36]. In this context, ensuring their reliable deployment is paramount to minimize undesirable risks. The reliability of VLMs goes beyond its most explored discriminative aspect [46,8,34] and requires providing robust uncertainty estimates. Nevertheless, this crucial factor has often been overlooked in the literature, with only a handful of recent works mainly

exploring the uncertainty of CLIP predictions from a calibration standpoint [43,25,38]. Despite their popularity [22,24], confidence calibration approaches lack theoretical guarantees of the actual model performance. In other words, these methods cannot estimate the most likely output and provide a verified probability of their predictive capabilities.

Conformal Prediction (CP) [1,40] is a machine learning framework that provides model agnostic, and *distribution-free*, finite-sample validity guarantees, ensuring reliability in predictive systems, and which has recently gained traction in modern neural networks [3,7]. In CP, the objective is to provide the user with prediction sets in which the probability of the correct label being included satisfies a user-specified coverage level, e.g., in the medical context, CP can ensure that the correct diagnosis (e.g., the presence of a given disease) is included in the predictive sets 95% of the time. *Split conformal prediction* (SCP) [26,40], a variant of CP techniques, is gaining increasing attention for modeling uncertainties in vision black-box predictors [3,7], i.e., models for which only access to their outputs is granted. SCP provides a practical scenario to incorporate such guarantees (marginalized over the entire test set) by leveraging a *calibration set*, which is assumed to be exchangeable w.r.t. test data [40]. Despite the promising performance of SCP in task-specific solutions, [3,18], its deployment in modern specialized VLMs remains less explored. Indeed, VLMs are strong zero-shot predictors [28], which can perform open-set black-box predictions on general concepts [28,23]. However, as recent evidence points out, its performance for such concepts highly relies on the pre-training concept and domain frequencies [39,21]. Note that, while this limitation does not affect the coverage guarantees of SCP, as these are distribution-free, it can affect the efficiency of the final conformal solution and, consequently, its practical utility. Transfer learning in VLMs has been widely studied in recent literature to mitigate such domain gap, where black-box Adapters, e.g., linear probe strategies, have been proposed [34,10] to adapt VLMs to downstream tasks by leveraging the embedding representations in the few-shot data regime. However, as we highlight in this work, this solution does not apply to the conformal prediction framework. In particular, adjusting these classifiers on the available small labeled calibration set might produce a covariate shift in the output probability distribution w.r.t. test data, breaking the exchangeability assumptions of CP. In this context, the theoretical guarantees of conformal prediction will not hold, producing unreliable predictive sets, as illustrated in Figure 1. This observation motivates the following question: *Can we enhance the performance of medical VLMs in SCP settings while maintaining the same coverage guarantees on incoming test data?*

The main contributions of this paper can be summarized as:

- We explore the SCP framework for medical VLMs, equipping them with trustworthiness guarantees in the popular few-shot scenario.
- To address transfer learning in VLMs, we introduce *transductive split conformal adaptation* (SCA-T). In this unsupervised framework, calibration and test data are jointly transferred to prevent breaking the exchangeability assumptions, thereby maintaining coverage guarantees.

- We proposed a transductive solver built upon well-established knowledge on information maximization, enabling the introduction of a specific label-marginal regularization estimated from the calibration set.
- Comprehensive experiments on 9 public datasets and 3 modality-specialized medical VLMs demonstrate the benefits of SCA-T to enhance average efficiency and conditional coverage over SCP, with up to +18% gains.

2 Related Work

VLMs transfer is typically carried out in the few-shot data regime, where Prompt Learning [46] and black-box Adapters [8,10,34] are popular strategies. These have been recently assessed in medical VLMs [32], showing on-par performance, despite Adapters being more efficient, as they operate over embedding representations. Concretely, SoTA Adapters are text-informed linear probes with text ensembles [10], or constraint optimization objectives [34]. However, these methods do not leverage the test data’s underlying distribution; that is, they are not transductive, which is our focus. The recent SoTA for transductive black-box adaptation in VLMs is TransCLIP [44], which combines well-established knowledge in unsupervised Gaussian Mixture Models with textual, zero-shot priors.

Conformal prediction. Current trends on CP for black-box image classifiers follow the SCP framework [3]. These works have focused on providing non-conformity measures to create conformal sets, e.g., by improving adaptability. Some non-conformity scores are LAC [31], which employs the raw probabilities; APS [30], accumulating the total uncertainty of a target category by accumulating the sorted softmax scores; or RAPS [3], which integrates penalties to promote smaller sets. Few works have explored SCP in medical image classification, e.g., by proposing ordinal non-conformity scores for grading tasks [19] or group-calibrated predictors [18], typically evaluated in specific applications. Note, however, that these models are intensively adjusted on a data corpus independent and identically distributed (i.i.d.) w.r.t. calibration/testing, an unrealistic scenario for pre-trained VLMs, which are our focus.

Conformal prediction in VLMs. Some very recent works have explored CP in VLMs. Conf-OT [35] proposes optimal transport to re-balance the zero-shot logits and address domain shift, and FCA [37] deploys a full-conformal adaptation pipeline for both adapting and conformalizing the outputs. The solution presented in this work is more flexible than Conf-OT [35], since we assume access to the embedding representations. Despite being less flexible than FCA [37], our solution enables adaptation without scaling compute with task complexity, i.e., number of classes, a key limitation of full-conformal scenarios.

3 Methods

3.1 Zero-shot models

Contrastive VLMs are trained on large heterogeneous datasets to encode similar representations between paired image and text information. Their visual and

text encoders project data points into an ℓ_2 -normalized 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 target classes, probabilities can be obtained as:

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

where $\mathbf{v}^\top \mathbf{w}$ is the dot product, and $\mathbf{p}(\mathbf{W}, \mathbf{v}) = (p_c(\mathbf{W}, \mathbf{v}))_{1 \leq c \leq C}$ corresponds to the temperature-scaled predicted probabilities vector.

Text-driven predictions. Contrastive VLMs enable zero-shot predictions, i.e., no need to explicitly learn $\mathbf{W}$, by embedding a textual description for each label into the weight matrix, i.e., $\mathbf{w}_c = \mathbf{t}_c$. Generally, the zero-shot prototypes are the average of the ℓ_2 -normalized text embeddings using a set of J text different templates, e.g., "A photo of a [CLS].", such that $\mathbf{t}_c = \frac{1}{J} \sum_{j=1}^J \mathbf{t}_{cj}$.

3.2 Conformal prediction

Preliminaries. Let us define an image classification task, where $(\mathbf{x}, y)$ denotes random data points. Also, we denote a black-box deep network, $\pi(\cdot)$, which outputs probability assignments, $\mathbf{p} = \pi(\mathbf{x})$, for a set of labels, $\mathcal{Y} = \{1, 2, \dots, C\}$.

Conformal prediction. In CP [40], the goal is to produce prediction sets that contain the true label with a user-specified probability. Formally, the aim is constructing a set-valued mapping $\mathcal{C} : \mathcal{X} \rightarrow 2^{\mathcal{C}}$, such that:

$$\mathcal{P}(Y \in \mathcal{C}(\mathbf{x})) \geq 1 - \alpha, \quad (2)$$

where $\alpha \in (0, 1)$ denotes the desired error rate, and $\mathcal{C}(\mathbf{x}) \subset \mathcal{Y}$ is the prediction set. Eq. 2 is known as the *coverage guarantee* [40], and is *marginal* over $\mathcal{X}\mathcal{Y}$.

Split conformal prediction. SCP [26] deploys coverage guarantees for any pre-trained black-box classifier [17]. This framework uniquely requires a calibration set consisting of N labeled data points, $\mathcal{D}_{\text{cal}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, drawn from a data distribution which is at least exchangeable [40] w.r.t. test data, $\mathcal{D}_{\text{test}} = \{(\mathbf{x}_i)\}_{i=N+1}^{N+M}$. The SCP procedure is as follows: *i) First* a non-conformity score $s_i = \mathcal{S}(\pi(\mathbf{x}_i), y_i)$ is defined, where s_i is a degree of uncertainty; *ii) Second*, the $1-\alpha$ quantile of the non-conformity score is determined from calibration data, which will serve as a confidence threshold to satisfy a given coverage:

$$\hat{s} = \inf \left[s : \frac{|\{i \in \{1, \dots, N\} : s_i \leq s\}|}{N} \geq \frac{\lceil (N+1)(1-\alpha) \rceil}{N} \right]; \quad (3)$$

iii) Third, non-conformity scores of each test data point and label are evaluated. The output sets are composed of the labels whose measure falls within $\hat{s}$:

$$\mathcal{C}(\mathbf{x}) = \{y \in \mathcal{Y} : \mathcal{S}(\pi(\mathbf{x}), y) \leq \hat{s}\}. \quad (4)$$

3.3 Transductive Split Conformal Adaptation (SCA-T)

Transfer learning. We aim to enhance conformal sets of pre-trained VLMs. A straightforward solution would involve leveraging $\mathcal{D}_{\text{cal}}$ supervision to adjust a linear probe, following relevant literature [8, 34]. Nonetheless, as shown in Fig. 1 (Adapt+SCP), modeling the output distribution using these labels would lead to a covariate shift, resulting in unsatisfactory coverage guarantees on the test data. Hence, we propose using an unsupervised learning objective, not relying directly on calibration labels. Additionally, such optimization is transductive, i.e., it is performed on both the joint calibration and test data, ensuring uniform transfer and reducing potential distribution shifts between the two subsets.

Transductive solver. Let us denote a calibration dataset composed of labeled embedding representations, $\{(\mathbf{v}_i, y_i)\}_{i=1}^N$, and a set of exchangeable test data points, $\{(\mathbf{v}_i)\}_{i=N+1}^{N+M}$. The objective is to learn a label assignment on the joint embedding dataset, $\{(\mathbf{v}_i)\}_{i=1}^{N+M}$, by learning a set of optimal class prototypes, $p_c(\mathbf{W}^*, \mathbf{v}_i)$, in an unsupervised manner. To solve this, we can leverage transductive objectives based on mutual information maximization, e.g., TIM [4]:

$$\min_{\mathbf{W}} \quad \frac{\lambda}{(N+M)} \sum_{i=1}^{N+M} \sum_{c=1}^C p_c(\mathbf{W}, \mathbf{v}_i) \log(p_c(\mathbf{W}, \mathbf{v}_i)) - \sum_{c=1}^C \hat{\mathbf{m}} \log(\hat{\mathbf{m}}) \quad (5)$$

where $\hat{\mathbf{m}} = \frac{1}{N+M} \sum_{i=1}^{N+M} \mathbf{p}(\mathbf{W}, \mathbf{v}_i)$ is the estimated label-marginal distribution, and λ is a blending hyper-parameter. The first term is the sample-conditional entropy, whose minimization aims to push the decision boundaries from dense regions. The second term refers to the label-marginal Shannon entropy, whose maximization encourages the marginal distribution of labels to be uniform, acting as a regularizer to avoid non-trivial solutions. The overall optimization objective in Eq. 5 can be solved, for example, by gradient descent optimizers.

One can notice that the regularizer in *TIM is biased toward uniform label distributions and does not generalize to imbalanced data regimes*, which are natural in medical tasks. In SCP, it is worth noting that we have access to the properties of the target data distribution, which can be inferred from calibration data. In particular, the marginal distribution is known and can be calculated from the one-hot encoded labels, $\mathbf{y}_i$, such that $\mathbf{m} = \frac{1}{N} \sum_{i=1}^N \mathbf{y}_i$. Hence, we propose a mutual information solver that can absorb this information. Concretely, we replace the Shannon entropy term with a more flexible Kullback–Leibler divergence, which regularizes the overall solution toward the observed priors:

$$\min_{\mathbf{W}} \quad \frac{\lambda}{(N+M)} \sum_{i=1}^{N+M} \sum_{c=1}^C p_c(\mathbf{W}, \mathbf{v}_i) \log(p_c(\mathbf{W}, \mathbf{v}_i)) + \sum_{c=1}^C \mathbf{m} \log \frac{\mathbf{m}}{\hat{\mathbf{m}}} \quad (6)$$

Conformal sets. The final step is producing the output sets using the new classifier’s scores, $\mathbf{p}(\mathbf{W}^*, \mathbf{v})$. First, calibration, $\{(\mathbf{p}(\mathbf{W}^*, \mathbf{v}_i)^\top, y_i)\}_{i=1}^N$, and test, $\{(\mathbf{p}(\mathbf{W}^*, \mathbf{v}_i))\}_{i=N+1}^{N+M}$, data is separated again. Second, vanilla SCP is followed as in Section 3.2, but using the updated classifier (from Eq. 6): *i*) non-conformity scores, $s_i = \mathcal{S}(\mathbf{p}(\mathbf{W}^*, \mathbf{v}_i), y_i)$, are generated; *ii*) the $1 - \alpha$ quantile is search as in Eq. 3, and *iii*) conformal sets are generated on test data following Eq. 4.

4 Experiments

4.1 Setup

Medical VLMs. Three major medical modalities are explored. **Histology:** CONCH [20] is used as a histology-specialized model, using a ViT-B/16 visual backbone. **Ophthalmology:** FLAIR [36] is selected. **CXR:** CONVIRT [45] pre-trained on MIMIC [14] is used. FLAIR and CONVIRT follow the same architectural design, composed of BioClinicalBERT [2] and ResNet-50 encoders.

Downstream tasks. **Histology:** colon in NCT-CRC [15], prostate Gleason grading in SICAPv2 [33], and SkinCancer [16] are evaluated. **Ophthalmology:** diabetic retinopathy grading using MESSIDOR [6], myopic maculopathy staging in MMAC [27], and disease classification in FIVES [13], are assessed. **CXR:** five CheXpert [12] categories from [42], fine-grained categories from NIH-LT [41,9], and multi-class pneumonia classification in COVID [5,29]. These datasets include grading and fine-grained classification tasks, with an average of nearly 8 classes.

Transfer learning. Our transductive solver is adjusted using the joint calibration and test data. The class weights are initialized using the text prototypes. Then, they are updated to minimize Eq. 6 via full-batch gradient descent during 100 iterations. ADAM is the optimizer, with a base learning rate of 0.01 and a cosine decay scheduler. Finally, λ is set to 1.0. *Note that all hyper-parameters are fixed across all tasks and data regimes.*

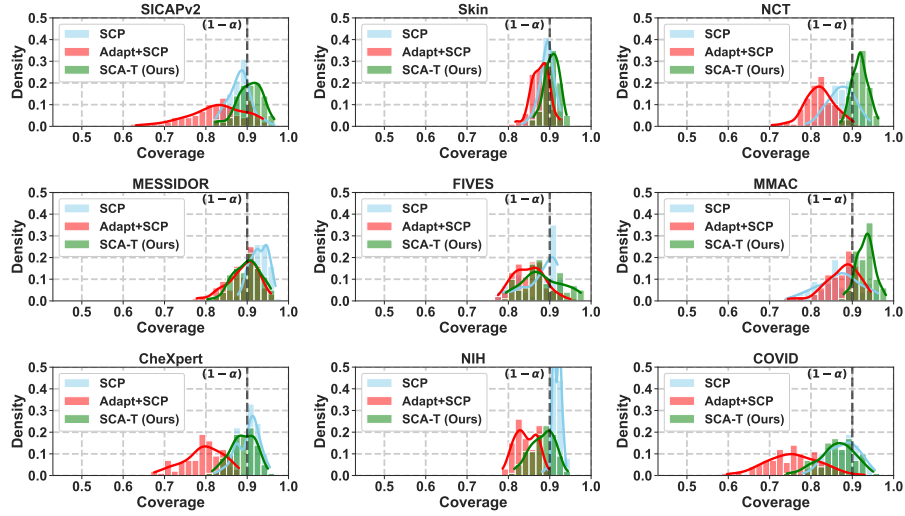
Baselines. Relevant black-box transductive baselines are included in our evaluation. First, the original TIM [4] is used. We select the gradient-based version, with similar training details as our method, maintaining the relative weight of the sample conditional entropy to 1, as it yields best performance. Second, we leverage the recent SoTA method for transduction using VLMs, i.e., TransCLIP [44].

Conformal prediction: Three popular non-conformity scores for image classification are assessed: LAC [31], APS [30], and RAPS [3], with RAPS’s hyper-parameters set to $k_{\text{reg}} = 1$ and $\lambda = 0.001$, which provided stable performance in [3]. Furthermore, prediction sets at error rates of $\alpha \in \{0.1, 0.05\}$ are generated.

Experimental protocol. **Calibration data:** $N = C \times K$ data points are sampled from training data following the label-marginal distribution of the downstream tasks, where $K = 16$ indicates the number of samples per class in the few-shot literature [32]. **Conformal strategies:** *i)* **SCP**, the standard split conformal prediction studied in vision classifiers [3,7] which operates over zero-shot predictions; *ii)* **Adapt+SCP**, which first adjusts a linear classifier and then follows the SCP over the new predictions, using the same calibration set for adaptation and calibration; and *iii)* **SCA-T**, the proposed transductive strategy that performs the transfer on the joint calibration and test data before conformal prediction. **Metrics** employed in conformal prediction literature [37] are computed: balance classwise accuracy (ACA), coverage (“Cov.”), average set size (“Size”), and class-conditioned coverage gap (“CCV”) [7]. Regarding computational efficiency, we assessed peak GPU memory usage in gigabytes (“GPU”) and inference time in seconds (“T”). The experiments are repeated 100 times with different sampling seeds.

Table 1: **Conformal prediction results**, averaged across modalities and tasks.

Method		$\alpha = 0.10$				$\alpha = 0.05$		
		ACA $\uparrow$	Cov.	Size $\downarrow$	CCV $\downarrow$	Cov.	Size $\downarrow$	CCV $\downarrow$
LAC	SCP	50.1	0.894	4.00	8.92	0.949	4.80	5.09
	SCA-T	55.2 _{+5.1}	0.898	3.30 _{-0.70}	7.47 _{-1.45}	0.952	4.03 _{-0.77}	4.03 _{-1.06}
APS	SCP	50.1	0.896	4.05	8.70	0.953	4.87	4.82
	SCA-T	55.2 _{+5.1}	0.900	3.35 _{-0.70}	7.18 _{-1.52}	0.954	4.08 _{-0.79}	3.97 _{-0.85}
RAPS	SCP	50.1	0.895	4.15	8.75	0.953	5.01	4.90
	SCA-T	55.2 _{+5.1}	0.903	3.40 _{-0.75}	7.50 _{-1.25}	0.954	4.18 _{-0.83}	4.57 _{-0.33}

Fig. 1: **Coverage analysis** per dataset and setting using LAC ($\alpha = 0.10$).

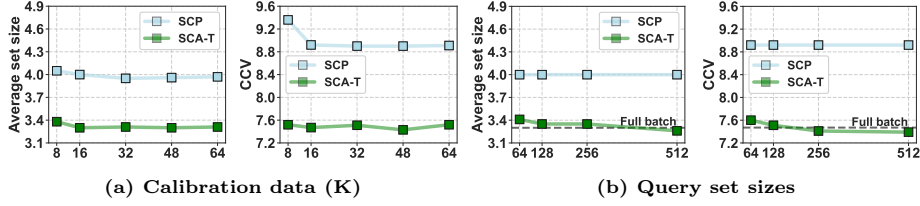
4.2 Results

Main results. Table 1 includes the CP results averaged across tasks. These demonstrate the effectiveness of the SCA-T. The proposed framework improves set efficiency by nearly 18% and adaptability (conditional coverage) by 7%–18%. The improvements are consistent across all non-conformity scores and error rates. It is worth noting that combining SCA-T with adaptive scores (e.g., APS) is particularly promising, especially if considering conditional coverage.

Coverage analysis. Fig. 1 shows the coverage distribution across random seed trials for each task and conformal strategy. One can notice that adapting and conforming the VLM outputs with the same calibration set (i.e., Adapt+SCP) breaks the exchangability of cal/test, and produces suboptimal coverage guarantees. In contrast, SCP provides nearly satisfactory coverage in most scenarios,

Table 2: **Transductive solvers** ($\alpha = 0.10$). **Red: unsatisfied error rates.**

Method	ACA $\uparrow$	T $\downarrow$	GPU $\downarrow$	Cov.	Size $\downarrow$	CCV $\downarrow$
LAC	50.1	0.00	-	0.894	4.00	8.92
TIM [4]	53.5 $\pm$ 3.4	1.12	0.6	0.888	3.96 $\pm$ 0.04	8.08 $\pm$ 0.84
TransCLIP [44]	54.8 $\pm$ 4.7	0.47	1.1	0.726	2.16 $\pm$ 1.84	22.31 $\pm$ 13.39
<i>Ours</i>	55.2 $\pm$ 5.1	1.04	0.6	0.898	3.30 $\pm$ 0.70	7.47 $\pm$ 1.45
APS	50.1	0.00	-	0.896	4.05	8.69
TIM [4]	53.5 $\pm$ 3.4	1.18	0.6	0.887	3.16 $\pm$ 0.89	7.47 $\pm$ 1.22
TransCLIP [44]	54.8 $\pm$ 4.7	0.40	1.1	0.733	2.52 $\pm$ 1.53	21.78 $\pm$ 13.09
<i>Ours</i>	55.2 $\pm$ 5.1	1.15	0.6	0.900	3.35 $\pm$ 0.70	7.18 $\pm$ 1.51

Fig. 2: **Detailed studies.** Results using LAC ($\alpha = 0.10$), averaged across tasks.

particularly considering finite-sample error. Furthermore, the proposed SCA-T consistently provides at least the same empirical coverage guarantees as SCP, often exceeding them, as in the NCT and MMAC datasets.

Baselines. Table 2 compares the proposed solver w.r.t. transductive methods. First, note that our solver provides the best accuracy increase (+5.1). Second, the baselines fail to meet the desired coverage, which showcases the importance of regularizing the marginal data distributions during transfer learning. Hence, even though TransCLIP produces smaller sets, those are unreliable. This evaluation also highlights the computational efficiency of SCA-T, which requires only an additional 0.8 seconds for inference and minimal computing resources.

Detailed analysis. Fig. 2 explores the limits of SCA-T on challenging data regimes. Results demonstrate consistent improvements over SCP for (a) small calibration sets, even in low-shot settings such as $K = 8$, and (b) processing test data in sequentially small test batches instead of operating fully transductive.

5 Discussion

We have explored split conformal prediction in medical VLMs as a tool for trustworthy image classification, and SCA-T as an efficient variant to improve the properties of the conformal sets. Despite its positive performance, some limitations remain. First, the performance of the entropy minimization-based solver in SCA-T relies on the quality of the initial VLM zero-shot predictions, which is

less flexible than our previous full-conformal adaptation scenario, FCA [37]. In scenarios where compute is not a limiting factor or when addressing only a few target categories, FCA could be a more appealing alternative. Second, CP provides exact theoretical guarantees under exchangeability, which can sometimes be challenging to achieve in healthcare. However, we observe that transductive solvers can mitigate such issues and warrant further research in the future.

Acknowledgments. This work was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC). We 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. Alex Gammerman, Volodya Vovk, V.V.: Learning by transduction. In: Conference on Uncertainty in Artificial Intelligence. pp. 148–156 (1998) [2](#)
2. Alsentzer, E., et al.: Publicly available clinical BERT embeddings. In: Clinical Natural Language Processing Workshop (2019) [6](#)
3. Angelopoulos, A.N., et al.: Uncertainty sets for image classifiers using conformal prediction. In: ICLR (2020) [2, 3, 6](#)
4. Boudiaf, M., et al.: Transductive information maximization for few-shot learning. In: NeurIPS. pp. 1–13 (2020) [5, 6, 8](#)
5. Chowdhury, M.E.H., et al.: Can ai help in screening viral and covid-19 pneumonia? IEEE Access **8**, 132665–132676 (2020) [6](#)
6. Decencière, E., et al.: Feedback on a publicly distributed image database: The messidor database. Image Analysis & Stereology **33**, 231–234 (07 2014) [6](#)
7. Ding, T., et al.: Class-conditional conformal prediction with many classes. In: NeurIPS. vol. 36 (2023) [2, 6](#)
8. Gao, P., et al.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision (2023) [1, 3, 5](#)
9. 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](#)
10. Huang, Y., et al.: Lp++: A surprisingly strong linear probe for few-shot clip. In: CVPR. pp. 23773–23782 (2024) [2, 3](#)
11. Huang, Z., et al.: A visual–language foundation model for pathology image analysis using medical twitter. Nature Medicine **29**, 1–10 (2023) [1](#)
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. Lei, J., et al.: Distribution-free predictive inference for regression. *Journal of the American Statistical Association* **113**(523), 1094–1111 (2018) [4](#)
18. Lu, C., et al.: Fair conformal predictors for applications in medical imaging. In: AAAI. vol. 36 (2022) [2](#), [3](#)
19. Lu, C., et al.: Improving trustworthiness of ai disease severity rating in medical imaging with ordinal conformal prediction sets. In: MICCAI. p. 545–554 (2022) [3](#)
20. Lu, M.Y., et al.: A visual-language foundation model for computational pathology. *Nature Medicine* **30**, 863–874 (2024) [1](#), [6](#)
21. Mayilvahanan, P., et al.: In search of forgotten domain generalization. In: ICLR (2025) [2](#)
22. Mehrtash, A., et al.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE TMI* **39**(12), 3868–3878 (2020) [2](#)
23. Moor, M., et al.: Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (4 2023) [2](#)
24. Murugesan, B., Silva-Rodríguez, J., Ayed, I.B., Dolz, J.: Class and region-adaptive constraints for network calibration. In: MICCAI (2024) [2](#)
25. Murugesan, B., et al.: Robust calibration of large vision-language adapters. In: ECCV (2021) [2](#)
26. Papadopoulos, H., et al.: Inductive confidence machines for regression. In: ECML. pp. 345–356 (2002) [2](#), [4](#)
27. Qian, B., et al.: A competition for the diagnosis of myopic maculopathy by artificial intelligence algorithms. *JAMA ophthalmology* **142**(11), 1006–1015 (2024) [6](#)
28. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021) [1](#), [2](#)
29. 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](#)
30. Romano, Y., et al.: Classification with valid and adaptive coverage. In: NeurIPS. vol. 33, pp. 3581–3591 (2020) [3](#), [6](#)
31. Sadinle, M., et al.: Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association* **114**(525), 223–234 (2019) [3](#), [6](#)
32. Shakeri, F., et al.: Few-shot adaptation of medical vision-language models. In: MICCAI. pp. 553–563 (2024) [1](#), [3](#), [6](#)
33. Silva-Rodríguez, et al.: 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) [6](#)
34. Silva-Rodríguez, J., et al.: A closer look at the few-shot adaptation of large vision-language models. In: CVPR. pp. 23681–23690 (2024) [1](#), [2](#), [3](#), [5](#)
35. Silva-Rodríguez, J., Ben Ayed, I., Dolz, J.: Conformal prediction for zero-shot models. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 19931–19941 (2025) [3](#)
36. Silva-Rodríguez, J., et al.: A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. *Medical Image Analysis* **99**, 103357 (2025) [1](#), [6](#)
37. Silva-Rodríguez, J., et al.: Full conformal adaptation of medical vision-language models. In: Information Processing in Medical Imaging (IPMI) (2025) [3](#), [6](#), [9](#)
38. Tu, W., et al.: An empirical study into what matters for calibrating vision-language models. In: ICML (2024) [2](#)
39. Udandarao, V., et al.: No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. In: NeurIPS (2024) [2](#)

- 40. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. Springer (01 2005) [2](#), [4](#)
- 41. 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](#)
- 42. Wang, Z., et al.: Medclip: Contrastive learning from unpaired medical images and text. In: EMNLP. pp. 1–12 (10 2022) [1](#), [6](#)
- 43. Yoon, H.S., et al.: C-TPT: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In: ICLR (2024) [2](#)
- 44. Zanella, M., et al.: Boosting vision-language models with transduction. In: NeurIPS (2024) [3](#), [6](#), [8](#)
- 45. Zhang, Y., et al.: Contrastive learning of medical visual representations from paired images and text. In: MHLC (2022) [1](#), [6](#)
- 46. Zhou, K., et al.: Learning to prompt for vision-language models. International Journal of Computer Vision (2022) [1](#), [3](#)