Localized Conformal Prediction for Image Classification with Vision-Language Models

Clément Fuchs*, Tim Bary*, Benoît Macq
ICTEAM, UCLouvain
Louvain-la-Neuve, Belgium

***Abstract*—Conformal predictions have attracted significant attention in the field of uncertainty quantification, mainly because of their strong marginal coverage guarantees. Full conditional guarantee is not an attainable goal, a well known fact in conformal predictions literature. As a result, several approaches have tried to approximate this behavior by adapting the conformal sets of test-time samples according to their similarity to calibration examples. Although the latter has gained traction and shown impressive performances for regression problems, its application to image classification remains under-explored. We conduct an extensive benchmarking on natural image classification tasks with vision-language models (VLMs), using our open source implementation of a recent localized conformal prediction algorithm. We show that straightforward usage of the cosine similarity between test-time and calibration visual features, an intuitive choice for VLMs, is not sufficient to improve over the non-local baselines. In response, we propose a simple non-linear transformation of the cosine similarities, which conserves marginal coverage guarantees and achieves statistically significant mean set sizes reduction. Code is available at github.com/cfuchs2023/lcp-vlm/.**

I. Introduction

Modern machine learning models achieve remarkable accuracy on visual tasks such as object detection, medical image analysis, and human action recognition. However, these models typically produce point predictions without quantifying uncertainty, which is a major shortcoming in safety-critical applications. In domains such as medical diagnostics, autonomous driving, or environmental monitoring, reliable uncertainty estimates help identify ambiguous inputs, flag unreliable predictions, and guide human oversight [4], [21].

Conformal prediction (CP) is a statistical framework for uncertainty quantification in machine learning. Given a predictive model and a set of calibration data, CP produces prediction sets or intervals that guarantee a user-defined level of coverage, assuming exchangeability of the data—a weaker assumption than i.i.d [23]. Due to its model-agnostic nature and theoretical guarantees, CP has been adopted in areas where reliable uncertainty estimates are crucial, such as medical imaging and human-in-the-loop systems.

Despite these strengths, a well-known limitation of CP is that it only guarantees marginal coverage (*i.e.*, averaged over the entire data distribution), but not conditional coverage for subgroups (*e.g.*, specific classes or feature regions) [1]. This leads to coverage disparities, where some subpopulations are under- or over-covered. These disparities often stem from the global nature of the calibration step, which assigns equal importance to all calibration examples.

Localized Conformal Prediction (LCP), introduced in [8], addresses this limitation by weighting calibration examples based on their similarity to the test sample. This yields test-specific prediction intervals that adapt to local distributional properties. While LCP retains the marginal coverage guarantees of standard CP, it can also offer improved local coverage under suitable conditions. So far, however, LCP has mainly been studied in regression settings and tested on synthetic datasets.

In this paper, we apply the LCP framework to classification tasks and benchmark it against standard CP on an extensive collection of datasets. Our approach integrates LCP into Vision-Language Models (VLMs) by leveraging their latent representations to define similarity. We show that using a non-linear transformation of the cosine similarity to weight the calibration set leads to improved set sizes and class-conditional coverage compared to traditional CP.

Our contributions are threefold:

1) **LCP for Classification:** We benchmark LCP for classification tasks using VLMs and evaluate it with different conformal classification scores such as LAC, APS, and RAPS. Additionally, we show that cosine similarity is a suboptimal weighting function for LCP in classification and propose a non-linear transformation to enhance its effectiveness.
2) **Extensive Benchmarking:** We evaluate our method on nine diverse image classification datasets, showing that LCP significantly improves average prediction set size.
3) **Open-Source Implementation:** We release the first, to our knowledge, open-source implementation of LCP for classification, compatible with a wide range of datasets and models.

The remainder of this article is structured as follows. Section II reviews CP and LCP, including classification-specific conformal scores and an overview of VLMs. Section III describes our experimental setup and introduces the proposed non-linear transformation of cosine similarity. In Section IV, we report and analyze the experimental results. Finally, Section V concludes the paper and outlines directions for future work.

*C. Fuchs and T. Bary contributed equally. C. Fuchs and T. Bary are funded by the MedReSyst project, FEDER and the Walloon Region. Computational resources have been provided by the CÉCI, funded by the F.R.S.-FNRS under Grant No. 2.5020.11 and the Walloon Region.

II. Background

A. Conformal prediction

Conformal predictions methods generally follow the same basic principle. First, conformal scores $s_i(k_i^*)$ are computed for n calibration samples, which differ from the training, validation, and test samples, with the knowledge of their ground-truth classes k_i^*. Then, the $1-\alpha$ quantile of these scores are computed as

$$q_\alpha = \text{Quantile}\left(\{s_i(k_i^*); 1 \leq i \leq n\}; \frac{n+1}{n}(1-\alpha)\right) \quad (1)$$

for a given error level α. Finally, for a test time sample x_{test}, all classes k such that $s_{\text{test}}(k) \leq q_\alpha$ are retained to form the conformal prediction set $C(\mathbf{x}_{\text{test}})$. One major difference between conformal prediction methods is the way they construct their conformal scores from the prediction vectors y_i, belonging to the $K-$simplex Δ_K, given by the model. In the following, we present four different scores which are ubiquitous in the literature.

a) TopK: The TopK conformal methods uses the rank of the class in the prediction vector as the conformal scores. Formally, let P_i the permutation function yielding the indexes sorting the components of $y_{i,k}$ in decreasing order. Then, the TopK conformal score reads as

$$s^{\text{TopK}}(k_i^*) = P_i(k_i^*). \quad (2)$$

b) Least Ambiguous set-valued Classifier: The Least Ambiguous set-valued Classifier (LAC) [12], [18], [23] uses conformal scores defined as follow :

$$s^{\text{LAC}}(k_i^*) = 1 - y_{i,k_i^*}. \quad (3)$$

Notice how this procedure may return empty sets for prediction vectors close to decision boundaries, a behavior considered undesirable in the literature.

c) Adaptive Prediction Sets: Subsequently, the Adaptive Prediction Sets (APS) method was introduced in [17]. It uses a different conformal score, which is the cumulated sum of the components of the soft label vector y_i, sorted in decreasing order, up to the component corresponding to the ground-truth label. Using the same notations as for TopK, let q^* the rank of the ground truth labels in the sorted prediction vector, i.e. the integer such that $P_i^{-1}(q^*) = k^*$. The APS score then reads

$$s^{\text{APS}}(k_i^*) = \sum_{q=1}^{q^*} y_{i,P_i^{-1}(q)}. \quad (4)$$

This procedure improves on the problem of non-empty conformal sets, but comes at the expense of prediction sets much larger than those constructed by the LAC method.

d) Regularized Adaptive Prediction Sets: As a result, regularized APS (RAPS) was introduced in [1]. It introduces two additional parameters, $k_{\text{reg}} \in \mathbb{N}$ and $\lambda_{\text{reg}} \in \mathbb{R}$, which penalize excessive set sizes. In detail, the APS score is modified as

$$s_i^{\text{RAPS}}(k^*) = s_i^{\text{APS}} + \lambda_{\text{reg}} \sum_{q=1}^{q^*} \mathbb{1}_{q > k_{\text{reg}}}. \quad (5)$$

The parameters k_{reg} and λ_{reg} are estimated from the data by splitting the calibration set into two additional subsets. One is dedicated to the computation of these parameters, the other to the computation of the quantile. We follow the same setup as [1] in our implementation.

B. Zero-Shot Prediction with Vision-Language Models

Generally, a VLM projects both images and textual descriptions into a common latent space, enabling measurement of their similarities. The textual descriptions of target classes provided by the user, so called textual prompts, are transformed into numerical tokens $\mathbf{c}_k$. The latter are then mapped by the textual encoder to normalized embeddings $\mathbf{t}_k$ on the unit-hypersphere of $\mathbb{R}^d$, where d is the latent space dimension. Similarly, the image $\mathbf{x}_i$ is processed by the visual encoder to produce embeddings $\boldsymbol{f}_i$ on the same unit-hypersphere. Prediction vectors are then obtained with a softmax transformation:

$$y_{i,k} = \frac{\exp(\mathbf{f}_i^T \mathbf{t}_k / T)}{\sum_j^K \exp(\mathbf{f}_i^T \mathbf{t}_k / T)} \quad (6)$$

where T is a parameter of the model, typically set to 0.01. The *Zero-Shot* prediction is then $\hat{k} = \text{argmax}_k \; y_{i,k}$. In this paper, we use models trained within the CLIP framework [16]. We investigate four different models, characterized by their visual encoder, either be ViT [5] based (ViT-B/16, ViT-L/14) or CNN [11] based (RN50, RN101). We use the prompts "a photo of a {class}." for all experiments. Notice how $\boldsymbol{f}_i^T \boldsymbol{f}_j$ is an intuitive choice to measure the similarity between images $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. The latter has been used successfully in the few-shot adaptation literature, for instance by Tip-Adapter [26].

C. Localized conformal prediction

We implement the algorithm presented in [8]. For a given test sample $\mathbf{x}_{\text{test}}$ and its corresponding projection in a latent space $\boldsymbol{f}_{\text{test}}$, LCP assigns greater importance to calibration examples that are more similar to it, using a localizer function $H(\boldsymbol{f}_{\text{test}}, \boldsymbol{f}_i) \in [0,1]$. This yields a test-specific, weighted empirical distribution over conformal scores, resulting in prediction sets that better adapt to local structure in the feature space.

In particular, given a conformal score s_i, $i = 1, \ldots, n$ for each calibration example, the weighted empirical distribution of scores is given by:

$$\hat{\mathcal{F}}(\boldsymbol{f}_{\text{test}}) = \sum_{i=1}^{n} \frac{H(\boldsymbol{f}_{\text{test}}, \boldsymbol{f}_i)}{\sum_{j=1}^{n} H(\boldsymbol{f}_{\text{test}}, \boldsymbol{f}_j)} \, \delta_{s_i}, \quad (7)$$

where δ_{s_i} is a point mass at the score s_i.

For classification with label space $\mathcal{Y} = \{1, \ldots, K\}$, a class k is included in the prediction set for the test point if its test-time score $s_{\text{test}}(k)$ does not exceed a quantile threshold:

$$C(x_{\text{test}}) = \{k \in \mathcal{Y} : s_{\text{test}}(k) \leq q_{\tilde{\alpha}}\}, \tag{8}$$

where $q_{\tilde{\alpha}}$ is the $1-\tilde{\alpha}$ quantile of the weighted distribution, and $\tilde{\alpha}$ is chosen to ensure a marginal coverage of $1-\alpha$. We refer to [8] for additional details regarding the algorithm and technical proofs about the coverage guarantee of LCP.

III. Methods

A. Experimental setting

a) Datasets: We follow the settings of previous works [27] and use 9 diverse images classification datasets: SUN397 [25] for classification of scenes, Aircraft [13] for aircrafts, EuroSAT [9] for satellite images, StanfordCars [10] for cars models, Food101 [2] for food items, Pets [15] for pet breeds, Flower102 [14] for flowers species, DTD [3] for textures and UCF101 [20] for actions recognition. We exclude Caltech101 [6] from the benchmark, as most backbones achieves an accuracy higher than the target coverage $1-\alpha$ for $\alpha = 0.1$. Details about the number of samples and labels distribution is provided in Supplementary Section A.

b) Calibration sets generation: We uniformly sample 1000 images from the dataset for calibration, and use the remaining data to measure the metrics exposed in the next Section. Therefore, the calibration have the same label distribution as the dataset (see Supplementary Section A). We use ten random seeds for each dataset, and use the same calibration / test splits for all methods.

c) Metrics: The CP literature usually wants to balance two objectives: (1) predictive efficiency (*i.e.*, producing small prediction sets); and (2) conditional validity (*i.e.*, maintaining the coverage guarantee across data subgroups).

Efficiency is measured by the *mean set size* over the set of test samples $\mathcal{S}_{\text{test}}$:

$$\text{Mean Set Size}(\mathcal{S}_{\text{test}}) = \frac{1}{|\mathcal{S}_{\text{test}}|} \sum_{\mathbf{x} \in \mathcal{S}_{\text{test}}} |C(\mathbf{x})|. \tag{9}$$

To assess conditional validity, we partition the test set by true class labels: $\mathcal{P} = \{\mathcal{S}_{\text{group}}\}$. For each group, the coverage Cov is:

$$\text{Cov}(\mathcal{S}_{\text{group}}) = \frac{1}{|\mathcal{S}_{\text{group}}|} \sum_{(\mathbf{x},y) \in \mathcal{S}_{\text{group}}} \mathbb{1}_{y \in C(\mathbf{x})}. \tag{10}$$

We follow previous works [7] and use two derived metrics:

- **CovGap:** average deviation from the target coverage:

$$\text{CovGap}(\mathcal{S}_{\text{test}}) = \frac{1}{|\mathcal{P}|} \sum_{\mathcal{S}_{\text{group}} \in \mathcal{P}} \left|\text{Cov}(\mathcal{S}_{\text{group}}) - (1-\alpha)\right|. \tag{11}$$

- **MCCC:** worst-case coverage across groups:

$$\text{MCCC}(\mathcal{S}_{\text{test}}) = \min_{\mathcal{S}_{\text{group}} \in \mathcal{P}} \text{Cov}(\mathcal{S}_{\text{group}}). \tag{12}$$

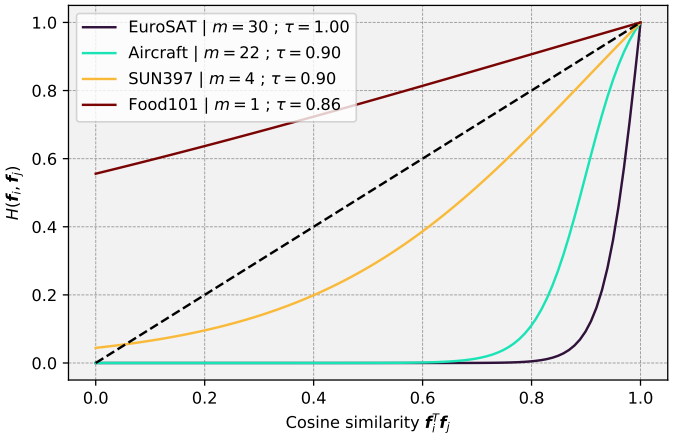

Fig. 1: Examples of optimal sigmoid transformations (see Eq. 13) obtained with our cross validation procedure (see Section III-B). The dotted black line corresponds to $H(\boldsymbol{f}_i, \boldsymbol{f}_j) = \boldsymbol{f}_i^T \boldsymbol{f}_j$, the naive version of the localized algorithm.

B. Transformation of Cosine Similarity

To define the localizer H in the LCP framework, we apply a normalized sigmoid transformation to the cosine similarity between samples:

$$H(\boldsymbol{f}_i, \boldsymbol{f}_j) = \frac{1 + \exp(-m(1-\tau))}{1 + \exp(-m(\boldsymbol{f}_i^T \boldsymbol{f}_j - \tau))}, \tag{13}$$

where $m > 0$ controls the sigmoid's steepness, and $\tau \in [0, 1]$ sets its inflection point. We tune m and τ via 5-fold cross-validation on the calibration set, so as to minimize the average set size. Finally, we take the median of the selected m values and the mean of the selected τ values across folds. In our experiments, we search over $\tau \in \{0.8, 0.9, 1.0\}$ and $m \in \{1, 2, \ldots, 30\}$.

IV. Results

A. Mean set sizes

We present results concerning the mean set sizes in Table I. To assess the significance of the differences in mean set sizes between local methods and non-local baselines, we perform a paired t-test [22] on the per-fold difference when it succeeds a Shapiro gaussianity test [19], and a Wilcoxon test [24] otherwise. We see that our approach achieves a decrease in mean set sizes for 9 out of 9 datasets for 3 out of 4 backbones compared to the non-local TopK, with statistical significance levels of least 0.01. Conversely, the naive approach fails to decrease the mean set sizes significantly compared to the non-local TopK, and even increases their sizes with statistical significance on several datasets and backbones (*e.g.* DTD with ViT-B/16, or UCF101 with RN50). This trend is confirmed with other non-local baselines, although not as sharply as with TopK. We see that the naive approach degrades mean set sizes with statistical significance for several other conformal scores, such as APS for Aircraft with ViT-B/16. However, our approach achieves a statistically significant decrease of the

mean set size in the latter setting, from 18.11 for the non-local baseline down to 17.43.

Overall, these results prove the effectiveness of our approach and the necessity of transforming the cosine similarities when using LCP. As an illustration, Figure 1 shows examples of sigmoidal transformations obtained with our cross validation strategy exposed in Section III-B.

B. Coverage metrics

We present results concerning coverage metrics in Table II, namely CovGap (see Eq. 11) and MCCC (see Eq. 12). Compared to the mean set sizes, it is more difficult to identify trends that hold across datasets and backbones. As a general observation, the local algorithms seem to more closely follow the behavior of the non-local baselines. However, there are still several instances where our approach improves MCCC, *e.g.*, for the LAC conformal score on Pets with ViT-B/16. However, on the same dataset with the same backbone, the MCCC is degraded compared to the non-local baseline when using the TopK conformal score. For the CovGap metric, the results obtained with the local algorithms are almost identical to the non-local baselines across all datasets and backbones. This result suggests that localization of conformal prediction is not sufficient to minimize the CovGap for image classification using VLMs.

V. Conclusion

In this paper, we presented an extensive benchmarking of localized conformal prediction on a variety of image classification tasks with several VLMs and conformal scores. We showed how the naive approach, using a cosine similarity in the visual latent space (an intuitive choice for VLMs), fails to improve over the non-local baselines, even degrading the mean set size in several cases. We thus introduced a sigmoidal transformation of the cosine similarities, using cross-validation on the calibration samples to tune its hyperparameters. This approach allowed to consistently improve over the non-local baselines across datasets and backbones, often achieving a statistically significant reduction of set sizes. Conversely, we showed that improvements in classical coverage-related metrics were less impacted by the local algorithms, with the coverage gap very closely following the behavior of the non-local baselines. We hope these findings as well as our open source code will contribute to the deployment of localized conformal prediction algorithms in real-world applications.

References

[1] A. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan, "Uncertainty sets for image classifiers using conformal prediction," *arXiv preprint arXiv:2009.14193*, 2020.

[2] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*. Springer, 2014, pp. 446–461.

[3] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3606–3613.

[4] J. C. Cresswell, Y. Sui, B. Kumar, and N. Vouitsis, "Conformal prediction sets improve human decision making," in *Forty-first International Conference on Machine Learning*, 2024.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[6] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *2004 conference on computer vision and pattern recognition workshop*. IEEE, 2004, pp. 178–178.

[7] L. Fillioux, J. Silva-Rodríguez, I. B. Ayed, P.-H. Cournède, M. Vakalopoulou, S. Christodoulidis, and J. Dolz, "Are foundation models for computer vision good conformal predictors?" *arXiv preprint arXiv:2412.06082*, 2024.

[8] L. Guan, "Localized conformal prediction: A generalized inference framework for conformal prediction," *Biometrika*, vol. 110, no. 1, pp. 33–50, 2023.

[9] P. Helber, B. Bischke, A. Dengel, and D. Borth, "Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 7, pp. 2217–2226, 2019.

[10] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 554–561.

[11] Y. LeCun, Y. Bengio *et al.*, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[12] J. Lei, A. Rinaldo, and L. Wasserman, "A conformal prediction approach to explore functional data," *Annals of Mathematics and Artificial Intelligence*, vol. 74, pp. 29–43, 2015.

[13] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.

[14] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian conference on computer vision, graphics & image processing*. IEEE, 2008, pp. 722–729.

[15] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3498–3505.

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[17] Y. Romano, M. Sesia, and E. Candes, "Classification with valid and adaptive coverage," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3581–3591, 2020.

[18] M. Sadinle, J. Lei, and L. Wasserman, "Least ambiguous set-valued classifiers with bounded error levels," *Journal of the American Statistical Association*, vol. 114, no. 525, pp. 223–234, 2019.

[19] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3-4, pp. 591–611, 1965.

[20] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[21] E. Straitouri and M. G. Rodriguez, "Designing decision support systems using counterfactual prediction sets," in *Forty-first International Conference on Machine Learning*, 2024.

[22] Student, "The probable error of a mean," *Biometrika*, pp. 1–25, 1908.

[23] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic learning in a random world*. Springer, 2005, vol. 29.

[24] F. Wilcoxon, "Individual comparisons by ranking methods," in *Breakthroughs in statistics: Methodology and distribution*. Springer, 1992, pp. 196–202.

[25] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 3485–3492.

[26] R. Zhang, W. Zhang, R. Fang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-adapter: Training-free adaption of clip for few-shot classification," in *European conference on computer vision*. Springer, 2022, pp. 493–510.

[27] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

TABLE I: Averaged set sizes (see Eq. 9) over 10 folds, for $\alpha = 0.1$. The best (*i.e.*, lowest) values are highlighted in **bold**. Statistical significance in the performance difference at the 0.05, 0.01 and 0.001 levels for a paired t-test (resp. Wilcoxon test) w.r.t. the non-local baseline are denoted by * (resp. °), ** (resp. °°), and *** (resp. °°°). L- denotes local approaches (see Section II-C). Naive corresponds to the black dotted line in Fig. 1, while the methods highlighted in pink corresponds to our procedure described in Section III-B. Standard deviation is noted after the ± signs.

(a) Results with ViT-B/16.

	UCF101	DTD	Pets	EuroSAT	StanfordCars	Flower102	Aircraft	SUN397	Food101
Zero-Shot Accuracy	67.5	43.3	89.1	48.3	65.6	70.8	24.9	62.6	85.9
TOPK	5.90 ±0.74	19.00 ±1.33	2.00 ±0.00	6.00 ±0.00	3.80 ±0.42	9.50 ±0.85	22.00 ±0.82	4.80 ±0.42	2.00 ±0.00
L-TOPK (naive)	6.23° ±0.63	19.98°°° ±1.10	1.19°°° ±0.21	5.31°°° ±0.19	3.80 ±0.42	8.27°°° ±0.80	22.36° ±0.82	4.50° ±0.44	2.00 ±0.00
L-TOPK	**4.72**°°° ±0.40	**16.71**°°° ±0.87	**1.17**°°° ±0.06	**4.28**°°° ±0.13	**3.26**°°° ±0.25	**7.57**°°° ±0.62	**20.10**°°° ±0.80	**4.38**°° ±0.37	**1.63**°°° ±0.17
LAC	3.42 ±0.25	15.76 ±1.24	**1.06** ±0.03	4.20 ±0.06	**2.46** ±0.11	5.50 ±0.66	18.11 ±0.96	3.47 ±0.39	**1.20** ±0.05
L-LAC (naive)	3.80°°° ±0.29	16.97°°° ±1.04	1.07° ±0.04	4.16°°° ±0.07	2.63°°° ±0.16	4.93°°° ±0.52	19.07°°° ±0.87	3.54°° ±0.37	1.29°°° ±0.07
L-LAC	**3.32**°° ±0.27	**14.62**°°° ±0.93	1.09° ±0.05	**3.94**°°° ±0.13	2.46 ±0.17	**4.92**°°° ±0.44	**17.43**° ±0.81	**3.44** ±0.35	1.41 ±0.62
APS	6.41 ±0.34	16.64 ±0.88	1.63 ±0.03	4.63 ±0.12	3.72 ±0.18	6.41 ±0.55	18.48 ±0.62	7.68 ±0.82	**2.32** ±0.11
L-APS (naive)	6.80°°° ±0.44	17.94°°° ±0.77	1.65°° ±0.03	4.58°°° ±0.12	3.89°°° ±0.18	**5.85**°°° ±0.44	19.31°°° ±0.56	7.77°° ±0.84	2.49°°° ±0.12
L-APS	**6.20**° ±0.41	**15.86**°°° ±0.73	**1.63** ±0.03	**4.52**°°° ±0.11	**3.70** ±0.19	6.15° ±0.49	**18.11**° ±0.52	**7.67** ±0.82	2.35 ±0.15
RAPS	5.10 ±1.01	15.47 ±2.74	1.65 ±0.04	4.52 ±0.44	3.55 ±0.33	5.88 ±1.89	15.77 ±2.01	**4.49** ±0.45	1.97 ±0.10
L-RAPS (naive)	5.36** ±1.13	16.01** ±2.60	1.67*** ±0.05	4.46 ±0.42	3.67°° ±0.42	**5.39** ±1.19	16.36** ±2.12	4.54 ±0.40	2.07*** ±0.15
L-RAPS	**4.85** ±0.57	**14.67*** ±2.32	**1.65** ±0.04	**4.07**** ±0.43	**3.53** ±0.35	5.59 ±1.33	**15.34*** ±1.64	4.50* ±0.45	**1.96** ±0.11

(b) Results with ViT-L/14.

	UCF101	DTD	Pets	EuroSAT	StanfordCars	Flower102	Aircraft	SUN397	Food101
Zero-Shot Accuracy	75.1	53.4	93.5	60.3	76.9	79.6	32.5	67.7	90.9
TOPK	3.40 ±0.52	12.70 ±0.67	**1.00** ±0.00	3.50 ±0.53	2.50 ±0.53	4.10 ±0.57	9.40 ±0.52	4.10 ±0.32	1.30 ±0.48
L-TOPK (naive)	3.20 ±0.35	13.48*** ±0.83	1.00 ±0.00	2.90°° ±0.15	2.41 ±0.44	3.70° ±0.47	9.56 ±0.49	3.95 ±0.09	1.30 ±0.48
L-TOPK	**2.63***** ±0.23	**11.68***** ±0.52	1.00 ±0.00	**2.79**°°° ±0.15	**1.99**°° ±0.17	**2.78***** ±0.19	**8.40***** ±0.36	**3.59***** ±0.26	**1.06** ±0.10
LAC	**1.94** ±0.09	9.91 ±0.60	**0.95** ±0.02	2.46 ±0.19	**1.57** ±0.08	2.03 ±0.22	7.40 ±0.31	2.92 ±0.27	**0.98** ±0.03
L-LAC (naive)	2.06*** ±0.06	10.89*** ±0.72	0.96*** ±0.02	**2.40***** ±0.18	1.64*** ±0.09	2.05 ±0.24	7.87*** ±0.40	2.97** ±0.26	1.02*** ±0.03
L-LAC	1.94 ±0.09	**9.54**** ±0.51	0.96 ±0.03	2.41 ±0.14	1.59 ±0.10	**1.97** ±0.15	**7.24** ±0.37	**2.89**** ±0.27	1.05 ±0.11
APS	3.81 ±0.19	11.09 ±0.55	1.33 ±0.01	3.23 ±0.12	**2.35** ±0.09	**3.04** ±0.12	8.74 ±0.34	7.27 ±0.57	**1.70** ±0.07
L-APS (naive)	3.90*** ±0.22	12.15*** ±0.73	1.35*** ±0.01	3.19°°° ±0.12	2.44*** ±0.11	3.04 ±0.12	9.22*** ±0.40	7.41 ±0.58	1.81*** ±0.08
L-APS	**3.71**** ±0.20	**10.86**** ±0.48	**1.33**** ±0.01	**3.11***** ±0.08	2.35 ±0.11	3.07 ±0.10	**8.57***** ±0.36	**7.21** ±0.48	1.74 ±0.10
RAPS	2.92 ±0.43	9.41 ±1.75	1.29 ±0.02	3.23 ±0.37	**2.25** ±0.08	2.88 ±0.28	7.87 ±0.51	3.80 ±0.12	**1.44** ±0.06
L-RAPS (naive)	2.98°° ±0.45	9.88*** ±1.82	1.31*** ±0.02	3.19** ±0.35	2.32*** ±0.09	2.89 ±0.26	8.14*** ±0.55	3.82* ±0.11	1.50°°° ±0.07
L-RAPS	**2.88** ±0.38	**9.22**** ±1.70	**1.29*** ±0.02	**3.04*** ±0.39	2.26 ±0.09	**2.80*** ±0.22	**7.68*** ±0.53	**3.79** ±0.11	1.45 ±0.07

(c) Results with RN50.

	UCF101	DTD	Pets	EuroSAT	StanfordCars	Flower102	Aircraft	SUN397	Food101
Zero-Shot Accuracy	61.9	42.8	85.7	36.1	55.8	66.0	17.0	58.8	77.4
TOPK	8.40 ±0.52	19.90 ±1.20	2.00 ±0.00	5.60 ±0.52	6.50 ±0.53	12.50 ±1.65	34.20 ±1.93	6.80 ±0.63	3.10 ±0.32
L-TOPK (naive)	8.82°° ±0.52	21.02*** ±1.43	2.00 ±0.00	4.62°°° ±0.49	6.49 ±0.55	11.14°°° ±1.69	34.53 ±1.93	**6.14***** ±0.61	3.10 ±0.32
L-TOPK	**7.42**°°° ±0.34	**18.84***** ±1.04	**1.52***** ±0.08	**4.58**°°° ±0.07	**5.66***** ±0.59	**11.01***** ±1.51	**30.66***** ±1.57	6.15*** ±0.63	**2.73***** ±0.24
LAC	5.52 ±0.35	17.52 ±0.97	**1.27** ±0.05	4.71 ±0.17	4.50 ±0.46	7.79 ±0.74	**28.15** ±1.57	4.82 ±0.33	**1.94** ±0.17
L-LAC (naive)	6.01*** ±0.39	18.63*** ±1.13	1.30*** ±0.05	4.56*** ±0.16	4.75*** ±0.44	**7.34***** ±0.85	28.82*** ±1.65	4.74*** ±0.36	2.19*** ±0.19
L-LAC	**5.37**** ±0.36	**16.91***** ±1.01	1.28 ±0.07	**4.40***** ±0.08	**4.49** ±0.53	7.60 ±0.88	28.40 ±1.67	**4.73***** ±0.36	2.01 ±0.21
APS	8.51 ±0.53	18.37 ±0.83	1.97 ±0.05	5.49 ±0.13	**6.43** ±0.41	8.80 ±0.70	29.74 ±1.47	10.16 ±0.98	**3.76** ±0.31
L-APS (naive)	8.76*** ±0.62	19.33*** ±0.96	2.01*** ±0.05	5.33*** ±0.12	6.74*** ±0.47	**8.56*** ±0.61	30.75*** ±1.63	**10.04*** ±0.97	4.01*** ±0.31
L-APS	**8.30**** ±0.58	**17.92**** ±0.82	**1.96** ±0.06	**5.05***** ±0.11	6.44 ±0.41	8.94 ±0.87	**29.56** ±1.31	10.05* ±0.98	3.76 ±0.33
RAPS	7.14 ±1.11	16.22 ±1.85	1.99 ±0.05	5.04 ±0.77	5.42 ±0.62	6.97 ±1.34	**27.58** ±3.31	6.26 ±0.72	3.09 ±0.38
L-RAPS (naive)	7.38** ±1.25	16.68** ±1.76	2.01°° ±0.05	4.94*** ±0.74	5.61°° ±0.70	**6.91** ±1.64	28.91 ±3.66	**6.24** ±0.74	3.34°° ±0.52
L-RAPS	**7.07** ±1.07	**15.45**** ±1.94	**1.98** ±0.06	**4.23***** ±0.58	**5.37*** ±0.63	7.01 ±1.32	27.72 ±3.31	6.25 ±0.76	**3.08** ±0.39

(d) Results with RN101.

	UCF101	DTD	Pets	EuroSAT	StanfordCars	Flower102	Aircraft	SUN397	Food101
Zero-Shot Accuracy	61.0	37.1	86.9	32.8	63.2	64.4	18.1	59.0	80.7
TOPK	6.80 ±0.63	19.00 ±0.82	2.00 ±0.00	8.10 ±0.32	4.50 ±0.53	10.10 ±0.88	30.30 ±1.49	6.70 ±0.82	3.10 ±0.32
L-TOPK (naive)	6.83 ±0.66	18.47° ±0.84	**1.00***** ±0.00	7.12°°° ±0.31	4.39 ±0.50	**8.09***** ±0.65	**27.97***** ±1.28	7.43°°° ±0.59	2.76° ±0.41
L-TOPK	**5.73***** ±0.33	**17.84***** ±0.90	1.50*** ±0.06	**6.69***** ±0.18	**3.89***** ±0.41	8.47*** ±0.77	28.90*** ±1.30	**6.21**** ±0.82	**2.62***** ±0.38
LAC	**4.50** ±0.39	17.05 ±0.90	1.21 ±0.03	7.08 ±0.10	**2.96** ±0.27	6.49 ±0.87	27.37 ±1.78	4.81 ±0.56	**1.76** ±0.16
L-LAC (naive)	4.82°°° ±0.44	16.67** ±1.00	**1.17***** ±0.03	6.98*** ±0.09	3.10*** ±0.25	**5.45**°°° ±0.72	**25.80***** ±1.70	5.41*** ±0.54	1.78 ±0.17
L-LAC	4.55 ±0.51	**16.36***** ±0.85	1.23* ±0.05	**6.64***** ±0.09	3.06 ±0.41	7.86*** ±0.65	26.55** ±1.71	**4.73*** ±0.49	1.83* ±0.19
APS	8.07 ±0.37	18.01 ±0.79	1.86 ±0.02	7.46 ±0.19	4.38 ±0.23	7.99 ±0.70	28.97 ±1.85	10.84 ±0.91	**3.36** ±0.15
L-APS (naive)	8.64*** ±0.39	17.57** ±0.96	**1.82***** ±0.03	7.34*** ±0.17	4.61*** ±0.23	**7.01***** ±0.54	**27.34***** ±1.81	12.44°°° ±0.91	3.45* ±0.17
L-APS	**7.75***** ±0.35	**17.41***** ±0.67	1.87 ±0.01	**6.96***** ±0.12	**4.38** ±0.23	7.86 ±0.65	28.24** ±2.07	**10.68** ±0.82	3.45 ±0.29
RAPS	5.39 ±0.38	15.25 ±2.13	1.85 ±0.10	5.87 ±0.89	3.95 ±0.35	7.11 ±2.27	25.58 ±4.23	6.16 ±0.92	2.81 ±0.22
L-RAPS (naive)	5.64*** ±0.48	15.29 ±1.97	**1.81***** ±0.09	5.71 ±0.90	4.06*** ±0.37	**6.21**° ±1.15	24.88** ±4.39	6.80** ±1.28	2.83 ±0.22
L-RAPS	**5.38** ±0.41	**15.17** ±1.94	1.83° ±0.10	**5.31***** ±1.02	**3.91** ±0.33	6.65 ±1.46	**24.85**° ±3.52	**6.14** ±0.88	**2.79** ±0.21

TABLE II: CovGap / MCCC (see Eq. 11 and Eq. 12) averaged over 10 folds, for $\alpha = 0.1$. For CovGap, lower values are better, while for MCCC, higher values are better. The best results are highlighted in **bold**. L- denotes local approaches (see Section II-C). Naive corresponds to the black dotted line in Fig. 1, while the methods highlighted in pink corresponds to our procedure described in Section III-B.

(a) Results with ViT-B/16.

	UCF101	DTD	Pets	EuroSAT	StanfordCars	Flower102	Aircraft	SUN397	Food101
Zero-Shot Accuracy	67.5	43.3	89.1	48.3	65.6	70.8	24.9	62.6	85.9
TOPK	0.12 / 0.03	0.12 / 0.13	0.09 / **0.44**	0.09 / 0.64	0.11 / 0.17	0.15 / 0.00	0.12 / **0.00**	**0.09** / **0.07**	0.05 / **0.54**
L-TOPK (naive)	**0.12** / 0.05	0.12 / 0.17	0.09 / 0.36	0.09 / 0.63	0.11 / 0.17	0.16 / 0.00	**0.11** / 0.00	0.09 / 0.06	0.05 / 0.54
L-TOPK	0.12 / **0.07**	**0.12** / **0.21**	**0.08** / 0.41	**0.07** / **0.76**	**0.10** / **0.18**	**0.15** / **0.02**	0.12 / 0.00	0.09 / 0.06	**0.05** / 0.51
LAC	0.12 / 0.01	0.13 / 0.10	0.09 / 0.11	0.07 / 0.74	0.11 / 0.09	0.16 / 0.00	0.13 / 0.00	0.09 / 0.06	0.05 / 0.44
L-LAC (naive)	**0.11** / 0.02	0.13 / 0.14	0.09 / 0.13	0.07 / 0.75	0.11 / **0.12**	0.16 / 0.00	0.13 / 0.00	**0.09** / **0.07**	**0.05** / **0.51**
L-LAC	0.12 / **0.02**	**0.12** / **0.19**	**0.08** / **0.29**	**0.07** / **0.76**	**0.10** / 0.10	**0.16** / **0.01**	**0.12** / **0.00**	0.09 / 0.06	0.05 / 0.45
APS	0.11 / 0.05	0.12 / 0.14	0.08 / 0.90	0.06 / **0.82**	0.09 / 0.21	0.15 / 0.00	0.13 / **0.00**	0.08 / 0.21	**0.07** / 0.72
L-APS (naive)	0.11 / 0.05	0.12 / 0.17	0.08 / **0.91**	0.06 / 0.82	0.09 / **0.26**	0.15 / 0.00	0.12 / 0.00	0.08 / **0.22**	0.07 / **0.75**
L-APS	**0.11** / **0.05**	**0.12** / **0.19**	**0.08** / 0.90	**0.06** / 0.82	**0.09** / 0.23	**0.14** / **0.00**	**0.12** / 0.00	**0.08** / 0.21	0.07 / 0.72
RAPS	0.11 / 0.04	0.13 / 0.12	0.09 / 0.69	0.07 / **0.81**	0.10 / 0.27	**0.16** / 0.00	0.13 / **0.00**	0.08 / **0.08**	0.06 / 0.72
L-RAPS (naive)	**0.11** / **0.06**	0.12 / **0.13**	0.09 / **0.70**	**0.07** / 0.80	0.10 / **0.29**	0.16 / 0.00	**0.13** / 0.00	**0.08** / 0.08	0.06 / **0.74**
L-RAPS	0.11 / 0.04	**0.12** / 0.11	**0.09** / 0.70	0.08 / 0.68	**0.10** / 0.27	0.16 / **0.00**	0.13 / 0.00	0.08 / 0.08	**0.06** / 0.72

(b) Results with ViT-L/14.

	UCF101	DTD	Pets	EuroSAT	StanfordCars	Flower102	Aircraft	SUN397	Food101
Zero-Shot Accuracy	75.1	53.4	93.5	60.3	76.9	79.6	32.5	67.7	90.9
TOPK	**0.11** / 0.08	0.10 / 0.08	**0.08** / **0.47**	**0.08** / **0.78**	0.11 / **0.16**	0.14 / 0.00	0.08 / 0.21	**0.09** / **0.07**	0.05 / **0.62**
L-TOPK (naive)	0.11 / 0.06	**0.10** / 0.12	0.08 / 0.47	0.09 / 0.70	**0.11** / 0.16	0.15 / 0.00	**0.08** / **0.22**	0.09 / 0.06	0.05 / 0.62
L-TOPK	0.11 / **0.10**	0.10 / **0.21**	0.08 / 0.47	0.08 / 0.74	0.11 / 0.13	**0.14** / **0.00**	0.08 / 0.20	0.09 / 0.05	**0.05** / 0.54
LAC	0.12 / 0.01	0.10 / 0.04	0.09 / 0.24	0.08 / 0.68	0.11 / 0.06	0.15 / 0.00	0.09 / 0.16	0.09 / 0.13	0.05 / 0.49
L-LAC (naive)	**0.11** / 0.02	**0.10** / 0.07	0.09 / 0.28	0.08 / 0.67	0.10 / 0.10	0.15 / 0.00	0.09 / 0.18	**0.09** / **0.14**	**0.05** / **0.58**
L-LAC	0.12 / **0.05**	0.10 / **0.07**	**0.08** / **0.32**	**0.07** / **0.73**	**0.10** / **0.11**	**0.14** / **0.00**	**0.09** / **0.20**	0.09 / 0.13	0.05 / 0.50
APS	0.10 / 0.10	0.09 / 0.07	0.09 / 0.92	0.07 / 0.85	0.09 / 0.42	0.13 / **0.00**	0.09 / 0.10	0.08 / 0.27	**0.07** / 0.90
L-APS (naive)	0.10 / 0.10	**0.09** / 0.09	0.09 / **0.92**	0.07 / 0.84	0.09 / **0.45**	0.13 / 0.00	0.09 / 0.13	0.08 / 0.28	0.08 / **0.92**
L-APS	**0.10** / **0.12**	0.09 / **0.11**	**0.09** / 0.92	**0.07** / **0.86**	**0.09** / 0.43	**0.12** / 0.00	**0.09** / **0.15**	**0.08** / **0.29**	0.07 / 0.90
RAPS	0.10 / 0.13	0.10 / 0.05	**0.08** / 0.88	0.07 / **0.84**	0.09 / 0.27	0.13 / **0.00**	0.09 / 0.11	0.08 / **0.15**	**0.06** / 0.88
L-RAPS (naive)	**0.10** / **0.13**	**0.10** / 0.06	0.08 / **0.88**	0.07 / 0.84	0.10 / **0.28**	0.13 / 0.00	**0.09** / **0.14**	**0.08** / 0.15	0.07 / **0.89**
L-RAPS	0.10 / 0.10	0.11 / **0.08**	0.08 / 0.88	**0.07** / 0.78	**0.09** / 0.28	**0.13** / 0.00	0.09 / 0.13	0.08 / 0.15	0.06 / 0.88

(c) Results with RN50.

	UCF101	DTD	Pets	EuroSAT	StanfordCars	Flower102	Aircraft	SUN397	Food101
Zero-Shot Accuracy	61.9	42.8	85.7	36.1	55.8	66.0	17.0	58.8	77.4
TOPK	0.12 / 0.02	0.12 / 0.03	0.07 / **0.48**	**0.08** / **0.81**	0.10 / **0.08**	0.14 / 0.00	0.11 / 0.00	**0.08** / **0.03**	**0.05** / **0.70**
L-TOPK (naive)	**0.11** / **0.03**	0.12 / 0.04	0.07 / 0.48	0.11 / 0.68	**0.10** / 0.08	0.15 / 0.00	0.11 / 0.00	0.08 / 0.03	0.05 / 0.70
L-TOPK	0.12 / 0.02	**0.11** / **0.14**	**0.06** / 0.48	0.08 / 0.76	0.10 / 0.08	**0.14** / **0.01**	**0.10** / **0.00**	0.08 / 0.03	0.05 / 0.68
LAC	0.12 / 0.02	0.12 / 0.09	0.07 / 0.35	0.09 / 0.70	0.10 / 0.08	0.14 / **0.00**	0.11 / 0.01	0.08 / 0.02	0.05 / 0.66
L-LAC (naive)	**0.11** / 0.03	0.11 / 0.13	0.07 / 0.39	0.09 / 0.68	0.10 / **0.09**	0.15 / 0.00	0.11 / 0.01	0.08 / 0.03	**0.05** / **0.70**
L-LAC	0.12 / **0.03**	**0.11** / **0.19**	**0.07** / **0.42**	**0.07** / **0.75**	**0.10** / 0.08	**0.14** / 0.00	**0.11** / **0.01**	**0.08** / **0.03**	0.05 / 0.66
APS	0.11 / 0.08	0.12 / 0.10	0.07 / 0.90	0.06 / **0.83**	0.09 / 0.19	0.14 / 0.00	0.11 / 0.00	0.08 / 0.09	0.05 / 0.81
L-APS (naive)	**0.11** / **0.08**	**0.11** / 0.13	0.07 / **0.90**	0.06 / 0.81	0.09 / **0.21**	0.14 / 0.00	**0.11** / **0.01**	0.08 / 0.09	0.06 / **0.83**
L-APS	0.11 / 0.08	0.11 / **0.18**	**0.07** / 0.90	**0.06** / 0.83	**0.09** / 0.19	**0.14** / **0.01**	0.11 / 0.00	**0.08** / **0.09**	**0.05** / 0.81
RAPS	0.11 / 0.04	0.12 / 0.05	0.07 / 0.81	**0.08** / **0.72**	0.09 / 0.12	0.15 / **0.00**	0.11 / 0.00	0.08 / **0.03**	0.05 / 0.75
L-RAPS (naive)	**0.11** / **0.04**	**0.12** / 0.06	0.08 / 0.81	0.09 / 0.71	**0.09** / **0.12**	0.15 / 0.00	**0.11** / 0.00	0.08 / 0.03	0.05 / **0.76**
L-RAPS	0.11 / 0.04	0.12 / **0.06**	**0.07** / **0.81**	0.12 / 0.52	0.09 / 0.11	**0.15** / 0.00	0.12 / **0.00**	**0.08** / 0.03	**0.05** / 0.75

(d) Results with RN101.

	UCF101	DTD	Pets	EuroSAT	StanfordCars	Flower102	Aircraft	SUN397	Food101
Zero-Shot Accuracy	61.0	37.1	86.9	32.8	63.2	64.4	18.1	59.0	80.7
TOPK	0.12 / 0.00	0.11 / 0.07	0.09 / 0.11	0.11 / 0.61	0.10 / **0.00**	0.14 / 0.00	**0.11** / 0.00	0.09 / 0.01	0.05 / **0.52**
L-TOPK (naive)	**0.12** / 0.00	0.11 / 0.06	0.12 / 0.03	0.13 / 0.48	0.10 / 0.00	0.15 / 0.00	0.12 / 0.00	**0.08** / **0.02**	0.06 / 0.50
L-TOPK	0.12 / **0.00**	**0.11** / **0.19**	**0.08** / **0.13**	**0.09** / **0.67**	**0.09** / 0.00	**0.14** / **0.01**	0.12 / **0.01**	0.09 / 0.01	**0.05** / 0.51
LAC	0.12 / 0.01	0.11 / 0.15	0.09 / 0.04	0.10 / **0.64**	0.10 / 0.00	0.15 / 0.00	0.13 / 0.00	0.09 / 0.01	0.05 / 0.50
L-LAC (naive)	**0.11** / 0.01	0.11 / 0.14	0.09 / 0.03	0.10 / 0.62	0.10 / 0.00	0.15 / 0.00	0.13 / 0.00	**0.08** / **0.02**	0.05 / 0.51
L-LAC	0.12 / **0.01**	**0.11** / **0.22**	**0.08** / **0.15**	**0.09** / 0.63	**0.10** / **0.01**	**0.13** / **0.03**	**0.12** / **0.01**	0.09 / 0.01	**0.05** / **0.53**
APS	0.10 / 0.02	0.11 / 0.19	0.09 / 0.64	0.09 / **0.73**	0.09 / **0.00**	0.14 / 0.02	0.13 / 0.00	0.08 / 0.07	**0.06** / 0.75
L-APS (naive)	0.10 / **0.03**	0.11 / 0.18	0.09 / 0.62	0.09 / 0.72	0.09 / 0.00	0.14 / 0.01	0.13 / 0.00	0.08 / **0.09**	0.06 / **0.76**
L-APS	**0.10** / 0.03	**0.11** / **0.28**	**0.09** / **0.68**	**0.08** / 0.72	**0.09** / 0.00	**0.13** / **0.03**	**0.12** / **0.00**	**0.08** / 0.07	0.06 / 0.75
RAPS	0.12 / **0.00**	0.12 / 0.08	0.09 / 0.37	**0.15** / **0.45**	0.09 / **0.00**	0.14 / **0.00**	**0.13** / **0.00**	**0.08** / 0.02	0.06 / 0.61
L-RAPS (naive)	**0.11** / 0.00	**0.12** / 0.08	0.09 / 0.37	0.15 / 0.43	0.09 / 0.00	0.15 / 0.00	0.13 / 0.00	0.08 / **0.02**	0.06 / **0.61**
L-RAPS	0.12 / 0.00	0.12 / **0.11**	**0.09** / **0.38**	0.18 / 0.39	**0.09** / 0.00	**0.14** / 0.00	0.14 / 0.00	0.08 / 0.02	**0.06** / 0.61

Localized Conformal Prediction for Image Classification with Vision-Language Models

Supplementary Materials

A. Dataset details

TABLE A-I: Additional information on the datasets.

Dataset name	Classes	Samples	Task
UCF101	101	13320	Action classification
DTD	47	5640	Texture classification
Pets	37	7349	Pet breed classification
EuroSAT	10	27000	Satellite image classification
StanfordCars	196	16185	Car model classification
Flower102	102	8189	Flower classification
Aircraft	100	10000	Aircraft model classification
SUN397	397	39700	Scene classification
Food101	101	101000	Food classification

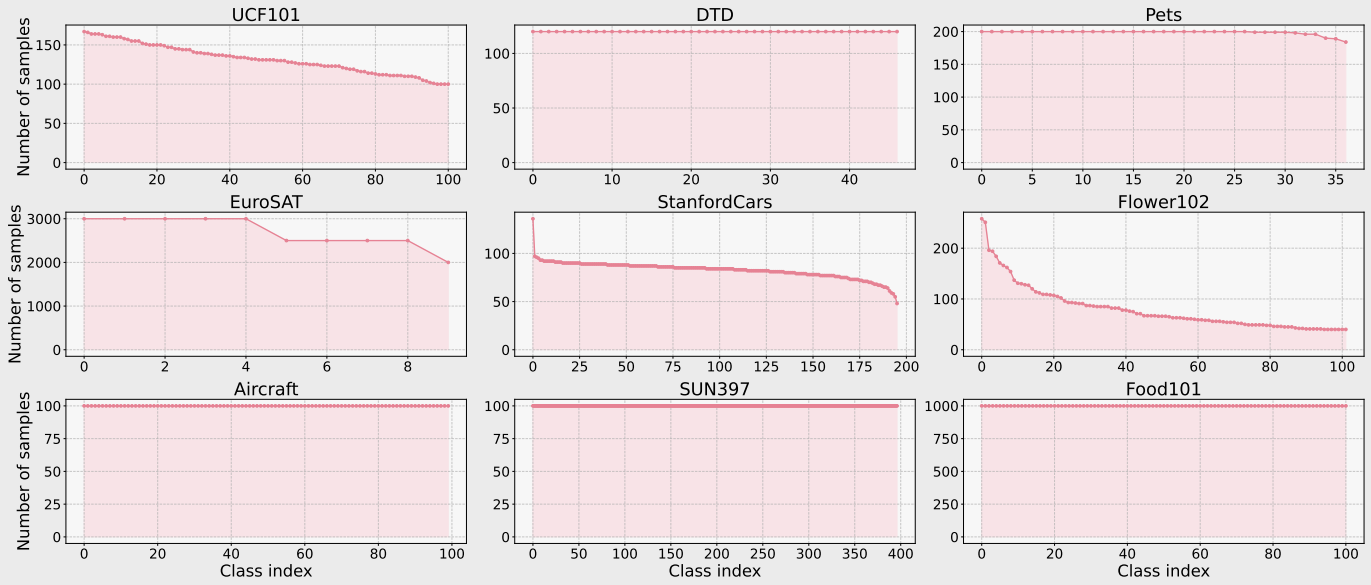

Fig. A-I: Labels distribution for each dataset. Note that our calibration sets are uniformly drawn from the dataset, therefore conserving the same distribution.