Conformal Prediction for Zero-Shot Models

Julio Silva-Rodríguez✉ Ismail Ben Ayed Jose Dolz
ÉTS Montréal
✉julio-jose.silva-rodriguez@etsmtl.ca

Abstract

Vision-language models pre-trained at large scale have shown unprecedented adaptability and generalization to downstream tasks. Although its discriminative potential has been widely explored, its reliability and uncertainty are still overlooked. In this work, we investigate the capabilities of CLIP models under the split conformal prediction paradigm, which provides theoretical guarantees to black-box models based on a small, labeled calibration set. In contrast to the main body of literature on conformal predictors in vision classifiers, foundation models exhibit a particular characteristic: they are pre-trained on a one-time basis on an inaccessible source domain, different from the transferred task. This domain drift negatively affects the efficiency of the conformal sets and poses additional challenges. To alleviate this issue, we propose Conf-OT, a transfer learning setting that operates transductive over the combined calibration and query sets. Solving an optimal transport problem, the proposed method bridges the domain gap between pre-training and adaptation without requiring additional data splits but still maintaining coverage guarantees. We comprehensively explore this conformal prediction strategy on a broad span of 15 datasets and three non-conformity scores. Conf-OT provides consistent relative improvements of up to 20% on set efficiency while being $\times 15$ faster than popular transductive approaches. We make the code available [1].

1. Introduction

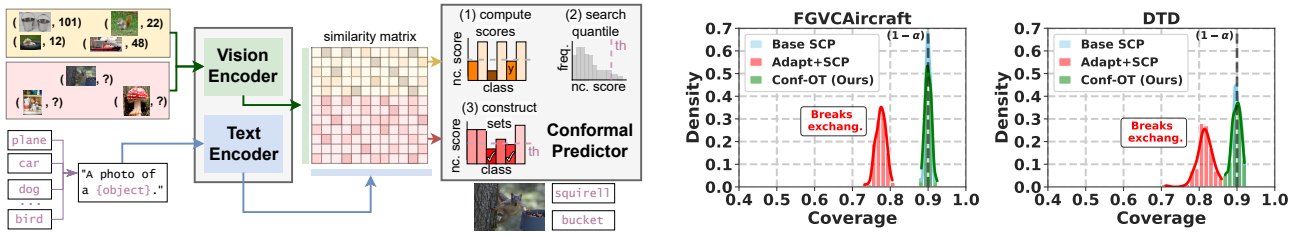
Deep learning is currently undergoing a paradigm shift with the emergence of large-scale vision-language models (VLMs), such as CLIP [49]. These models, which are trained on a massive amount of paired language and image data leveraging contrastive learning techniques, have demonstrated unprecedented zero-shot capabilities on a wide array of downstream visual tasks, including classification [44, 49], object detection [33, 41], segmentation [31, 36] or image synthesis [52], among many others. Inspired by the transferability power of VLMs, many efforts have focused on improving the discriminative performance of CLIP during adaptation to downstream tasks [18, 20, 26, 34, 56, 72, 74, 75].

Following their remarkable performance on general computer vision tasks, VLMs, and more particularly CLIP, are becoming increasingly popular in safety-critical scenarios, such as autonomous driving and medical imaging [32, 35, 54, 57]. Therefore, ensuring the reliability of model predictions is paramount for the safe deployment of these models in real-world applications, particularly considering their increasing adoption. Nevertheless, this crucial aspect has often been overlooked in the literature, with only a handful of recent works exploring the uncertainty of CLIP predictions from a calibration standpoint [42, 45, 55, 60].

Albeit popular, confidence calibration methods lack theoretical guarantees of the actual model performance. For example, these cannot estimate the most likely output (or set of outputs) and provide a verified probability of such prediction being correct. A principled solution is to quantify the uncertainty via conformal prediction (CP) frameworks [1, 46, 48, 53, 66], which has experienced a growing interest in more traditional machine learning models. CP provides confidence guarantees by yielding prediction sets that contain the correct label with a desired coverage level, e.g., they can ensure that the true category will be part of the predictive sets, on average, 95% of the time. Particularly, *split conformal prediction* [46, 66] provides a practical scenario to incorporate such marginal guarantees to *black-box* models by leveraging a small *calibration set*, which is assumed to be, at least, exchangeable with respect to test data [66]. With the growing interest in the trustworthiness of machine learning systems, many works have explored CP on classical image classification benchmarks [2, 13, 25, 59], including ImageNet [11] or CIFAR [28] datasets. For example, these works have focused on proposing novel criteria to create the predictive sets (non-conformity scores in conformal prediction literature) with improved efficiency [39], adaptiveness [2, 51], or conditional coverage [13].

Building on these observations, this work explores how

[1] https://github.com/jusiro/CLIP-Conformal



(a) Conformal prediction in VLMs

(b) Transfer learning and exchangeability.

Figure 1. **How to transfer black-box VLMs without breaking exchangeability?** In this work, we explore split conformal prediction (SCP) for VLMs (*see* (a)) to provide trustworthiness guarantees. These zero-shot models typically undergo adaptation to enhance their performance. However, leveraging the SCP calibration data for adaptation breaks the exchangeability assumption [66], which produces miss-coverage during inference (*see* (b)). We propose a transductive, unsupervised transfer to overcome such challenges, coined Conf-OT.

conformal prediction can be integrated into vision-language models to enhance their reliability while maintaining a competitive performance. Indeed, vision-language foundation models are promising *black-box* predictors, as evidenced by the existing literature [40, 49]. Nevertheless, their zero-shot predictive performance depends on the source data distribution and concept frequency [61]. Thus, they usually require an adaptation stage when severe domain gaps exist in the target tasks w.r.t. the pre-training data assembly. This adaptation can be performed with efficient linear probing solutions [34, 56]. However, this situation is problematic in the conformal prediction framework. In particular, if adjusting these classifiers using a few labeled examples, e.g., in a calibration set gathered for conformal prediction, the testing data scores may not be exchangeable w.r.t. calibration. Hence, the theoretical guarantees of conformal prediction will not hold, as illustrated in Fig. 1(b). This motivates the following question: *can the performance of VLMs in conformal settings be improved via transfer learning without additional data sources beyond the calibration set?*[2]

The main contributions of this paper can be summarized as:

- We introduce the split conformal prediction framework for large-scale pre-trained vision-language models, providing trustworthiness guarantees on the zero-shot predictions based on a small labeled calibration set.
- In contrast to the main corpus of recent literature in computer vision, which explores the CP framework using specialized models, VLMs are pre-trained on a generalist, inaccessible source domain, different from the downstream task and data distribution. To address this challenge, we propose Conf-OT, an unsupervised transfer learning framework that reduces the domain gap while maintaining coverage guarantees. Concretely, the proposed transductive strategy aims to solve the optimal transport problem on the joint calibration and query text-driven similarity matrix, producing a code assignment that respects the marginal properties of the target distribution.
- We provide extensive experiments to assess the performance of popular non-conformity scores atop black-box predictions produced by CLIP models, including 15 popular image classification benchmarks. The results demonstrate the effectiveness of Conf-OT to improve the set size efficiency and class-conditional coverage.
- Notably, upon the standard black-box conformal prediction paradigm, Conf-OT substantially outperforms recent transductive methods in the literature — even in the discriminative aspect — yet being a *training-free* approach, which requires minimal computational overhead.

[2]An additional labeled few-shot adaptation set could be introduced, thus keeping calibration data exchangeable to future queries. Nevertheless, demanding more labeled data might be *unrealistic* in practice, e.g., in critical scenarios such as detecting rare, low-prevalence diseases [17, 30, 54].

2. Related work

Zero-shot and transfer learning in VLMs. Contrastive VLMs such as CLIP models exhibit outstanding generalization capabilities [49], and enable zero-shot image classification without adaptation [40], despite being notably more accurate when concepts are represented during pre-training [61]. The latter limitation has directed the community toward developing data-efficient adaptation techniques, usually under the few-shot paradigm [18, 20, 26, 34, 56, 74, 75]. Particularly, efficient black-box Adapters, which only require embedded representations [18, 26, 34, 56, 74], are playing a pivotal role in this topic. The best results are obtained through advanced linear probing techniques that combine text-driven class prototypes with few-shot visual information [26, 34, 56]. As stated earlier, the reliability of these models remains less explored, with just a few recent works assessing the calibration aspect of VLMs [42, 55]. In contrast, our work focuses on a more principled framework for uncertainty quantification of VLMs outputs. To the best of our knowledge, there has been limited exploration of predictive uncertainty in vision-language models from a conformal prediction standpoint.

Transductive adaptation for image classification. A direction to improve pre-trained models consists of leverag-

ing the shared information of unlabeled test data, so-called *transduction* — in contrast to its more extended *inductive* counterpart, which makes independent predictions for each new data point. The first setting usually reports notable performance gains over the second [5], at the cost of additional test-time computation. Several transductive methods fine-tune the whole encoder [12, 67], whereas others promote lightweight black-box adaptation [5, 38, 73, 76]. The latter usually adjusts the class-wise prototypes in the feature space by exploiting mutual information on the query set [5, 62], or optimal transport [76]. Regarding VLMs, its transductive adaptation has been less explored, with only a few recent works [38, 73] focusing on the discriminative aspect. For example, [38] develops a solution for small tasks modeling the target data through a Dirichlet probability distribution, while TransCLIP [73] integrates a KL-divergence into a GMM clustering that encourages the predictions not to deviate from the textual prototypes.

Conformal prediction in vision classifiers. Conformal prediction is a framework for uncertainty quantification that produces statistically valid predictive regions [1, 46, 48, 53, 64]. This work focuses on *split conformal prediction* [46, 66], a resource-efficient, practical setting that allows conformalizing any black-box classifier. In particular, given a trained model that outputs logit predictions, it assumes access to a fresh labeled calibration set exchangeable [66] with testing data. This data is exploited to find a confidence-specific threshold from a non-conformity score, which is later employed for creating predictive sets with theoretical guarantees over such confidence level. To this end, different scores have been proposed [2, 39, 51]. Least Ambiguous Classifier, a.k.a. LAC [39] creates predictive sets by directly using the raw class probabilities. Adaptive Prediction Sets (APS) [51] computes the score by accumulating the sorted softmax values in descending order, and its regularized extension RAPS [2] tames the tail by enforcing small sets integrating explicit penalties.

Prior art on split conformal prediction has been validated on vision classification tasks [2, 9, 15, 51, 59, 69]. Nevertheless, these evaluations assume narrow scenarios, using specialized (only-vision) models, usually trained with a large corpus of data in-distribution w.r.t. calibration/test. This focus significantly differs from the current emerging paradigm in vision, driven by large-scale pre-training using VLMs, which are transferred to a broad corpus of downstream tasks [18, 49, 75]. Note that this zero-shot setting does not affect the coverage guarantees of split conformal prediction, which are distribution-free, but might hamper the efficiency and, hence, the usability of the produced sets.

Transduction in conformal prediction has been classically linked to *full conformal prediction* [1, 48, 53, 65]. However, this framework — see Appendix A — differs from the split setting addressed in our work. Particularly, it does not consider access to a calibration set. Instead, it evaluates each test data-label pair conformity by resorting to multiple model fits. It is worth noting that leveraging test data distribution is not exclusive to full conformal methods. For example, [19] explores a transfer learning scenario with a domain shift between train and calibration/test under the split conformal prediction umbrella. Particularly, the authors study a transductive strategy to reduce the domain gap during training, from which we draw inspiration. Nevertheless, the scenario in [19] assumes access to the source training data and requires training the base model, which drastically differs from our focus on foundation models.

3. Background

3.1. Zero-shot models

Contrastive vision-language pre-training. Large-scale VLMs, such as CLIP [49], are trained on large heterogeneous datasets to encode similar representations between paired image and text information. CLIP comprises a vision encoder, $f_\theta(\cdot)$, and a text encoder, $f_\phi(\cdot)$. These encoders 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.

Zero-shot inference. For a particular image classification task, CLIP-based models can provide predictions based on the similarity between category prompts, i.e., text descriptions for the new categories, and testing images. Given a set of K classes and an ensemble of J text prompts for each one, $\{\{\mathbf{t}_{kj}\}_{j=1}^{J}\}_{k=1}^{K}$, a common practice is to obtain a zero-shot prototype for each target category by computing the center of the ℓ_2-normalized text embeddings for each class, $\mathbf{t}_k = \frac{1}{J}\sum_{j=1}^{J}\mathbf{t}_{kj}$. Thus, for a given query image, the zero-shot prediction, $\hat{\mathbf{p}} = (\hat{p}_k)_{1\leq k\leq K}$, is obtained from the softmax cosine similarity between its vision embedding, $\mathbf{v}$, and category prototypes $\mathbf{t}_k$:

$$\hat{p}_k = \frac{\exp(\mathbf{v}^\top \mathbf{t}_k/\tau^{\text{CLIP}})}{\sum_{i=1}^{K}\exp(\mathbf{v}_i^\top \mathbf{t}_i/\tau^{\text{CLIP}})}, \tag{1}$$

where $l_k = (\mathbf{v}^\top \mathbf{t}_k/\tau^{\text{CLIP}})$ are the logits. Note that $\mathbf{v}^\top\mathbf{t}$ is the dot product, equivalent to cosine similarity, as vectors are ℓ_2-normalized. Thus, logits are similarity measures for each sample to the textual class prototypes scaled with τ^{CLIP}, a temperature parameter learned during the pre-training.

3.2. Conformal prediction

Preliminaries. Let us denote the black-box scores for an input image space of a zero-shot model, e.g., CLIP outputs in Eq. (1), as $\mathcal{X} \subset \mathbb{R}^{1\times K}$. Also, we denote their corresponding label space, $\mathcal{Y} = \{1, 2, ..., K\}$, and $(\mathbf{x}, y)$ as a random data pair sampled from a joint distribution $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$.

Split conformal inference. To provide trustworthiness on the outputs of a machine learning model, conformal pre-

diction [66] aims to produce predictive sets containing the ground truth label with a user-specified probability. Formally, the goal is to construct a set-valued mapping $\mathcal{C} : \mathcal{X} \rightarrow 2^K$, from a model output such that:

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

where $\alpha \in (0, 1)$ denotes the desired error rate (e.g., 10%), and $C(\mathbf{x}) \subset \mathcal{Y}$ is the prediction set. This is denoted as the *coverage guarantee*, and is *marginal* over $\mathcal{X}\mathcal{Y}$.

Split conformal prediction [46] assumes a practical setting for black-box models, enabling deploying coverage guarantees for any predictor [29]. First, it grants access to a labeled calibration subset $\mathcal{D}_{\text{cal}}= \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$. Second, it assumes that the test data, $\mathcal{D}_{\text{test}}= \{(\mathbf{x}_i)\}_{i=N+1}^{N+M}$, are i.i.d. or exchangeable [66] fresh data points, not used for training. *First*, the split conformal prediction process starts by defining a non-conformity score $s_i = \mathcal{S}(\mathbf{x}_i, y_i)$ for each calibration sample, where s_i is a measure of deviation between an example and the training data, which we will specify later. *Second*, the 1-α quantile of the non-conformity score is determined from calibration data, which will serve as a confidence threshold to satisfy a given coverage:

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

Third, for each testing sample, the non-conformity score for each label is calculated. The prediction set comprises labels whose non-conformity score falls within $\hat{s}$:

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

Non-conformity scores. Different criteria have been proposed, aiming to produce small (*a.k.a. efficient*) sets but able to model *adaptiveness*, e.g., larger predictive for uncertain test points. For the first, LAC [39] tends to produce the smallest possible predictive sets, while for the latter, adaptive scores such as APS [51] and RAPS [2] are popular options in vision. These are introduced in Appendix B.

Black-box setting. The standard split conformal prediction setting deployed in vision tasks [2, 13, 69] usually takes as input the raw logits produced by the base model, $\boldsymbol{l}_i$, and contemplates the possibility of controlling the sharpness of its distribution, e.g., by using temperature scaling [2, 69] before producing softmax scores, i.e., $\mathbf{p}_i = \sigma_k(\boldsymbol{l}_i/\tau)$, being $\sigma(\cdot)_k$ the softmax activation, and τ a temperature parameter. Once the classwise probabilities are obtained, these are used as input for computing non-conformity scores, i.e., $\mathbf{x}_i = \mathbf{p}_i$.

4. Proposed solution

4.1. Conformal prediction in zero-shot models

Motivation. Prior art in conformal prediction for vision [2, 13, 25, 59] assumes access to specialized black-box models pre-trained on a training subset drawn from the same data distribution as calibration and test. Nevertheless, this scenario is *unrealistic* when transferring cutting-edge foundation models, particularly zero-shot VLMs. These are pre-trained on multiple tasks from inaccessible source data that differs from the target domain.

Transfer learning setting. Let us assume a scenario in which a black-box model from a source domain, $\mathcal{D}_{\text{train}}$, produces logits for a set of target categories. Also, for a new task, there exists a labeled calibration set $\mathcal{D}_{\text{cal}}= \{(\boldsymbol{l}_i, y_i)\}_{i=1}^{N}$, and unlabeled testing data, $\mathcal{D}_{\text{test}}= \{(\boldsymbol{l}_i)\}_{i=N+1}^{N+M}$, and we aim to create conformal predictive sets. Importantly, $\mathcal{D}_{\text{cal}}$ and $\mathcal{D}_{\text{test}}$ are exchangeable distributions from a target domain, which are different from $\mathcal{D}_{\text{train}}$.

Problem statement. The first measure to produce efficient sets is to learn a transfer function from the source to the target domains. One naive option would involve leveraging $\mathcal{D}_{\text{cal}}$ supervision to adapt the black-box outputs, e.g., following few-shot adaptation literature [18, 56]. Nonetheless, it is crucial to consider the final conformal predictive scenario. As shown in Fig. 1(b), modeling the logit distribution to maximize the likelihood of label assignment using such labels would break the exchangeability assumptions.

4.2. Transductive conformal prediction

We propose a transfer learning strategy, which is: *i*) *unsupervised*, i.e., does not directly rely on label supervision, and *ii*) *transductive*, i.e., calibration and test (queries in the transductive literature) data points are jointly transferred. Thus, the proposed setting avoids introducing any distributional shifts that could potentially break the exchangeability assumption required in conformal prediction.

Optimal transport for transfer learning. We leverage well-established knowledge in optimal transport (OT) [10, 63] to learn a joint mapping from source to target domain in the label assignments in an unsupervised manner. Such technical choice is motivated by the capabilities of OT to produce assignments that respect, for instance, a given label-marginal distribution — estimated from the calibration set — thus reducing potential domain drifts. Our approach, coined **Conf-OT**, is detailed in Algorithm 1, and we describe each component below.

Learning objective. Let us consider the combined calibration and test sets, which are integrated into a similarity matrix, $\mathbf{S} \in \mathbb{R}^{K \times (N+M)}$. Each column represents the similarities to each category prototype, i.e., the logits of a given sample, $\boldsymbol{l}_i$, extracted as detailed in Sec. 3.1. Our goal is to find the joint probabilities matrix, $\mathbf{Q} \in \mathbb{R}_{+}^{K \times (N+M)}$, typically referred to as *codes*, which maximizes the similarity assignment. Note that each column in $\mathbf{Q}$, i.e., $\mathbf{q}_i$, is the prototype assignment for each sample. To achieve this, we propose to cast this task as an *optimal transport problem*, introducing marginal constraints for the expected target dis-

tribution. Formally, the search problem can be defined as:

$$\max_{\mathbf{Q}\in\mathcal{Q}} tr(\mathbf{Q}^\top \mathbf{S}), \tag{5}$$

where the matrix $\mathbf{Q}$ is relaxed to be an element of the *transportation polytope*:

$$\mathcal{Q} = \{\mathbf{Q} \mid \mathbf{Q}\mathbf{1}_{(N+M)} = \mathbf{m}, \mathbf{Q}^\top \mathbf{1}_K = \mathbf{u}_{(\mathbf{N+M})}\}, \tag{6}$$

such that $\mathbf{1}_{(\cdot)}$ denotes a column vector of ones, and $\mathbf{u}_{(\cdot)}$ an uniform distribution, being $(\cdot)$ the input vector length. In this element, $\mathbf{m}$ and $\mathbf{u}_{(\mathbf{N+M})}$ determine the marginal distributions expected in the target domain. First, $\mathbf{u}_{(\mathbf{N+M})} = \frac{1}{(N+M)}\mathbf{1}_{(N+M)}$, is the sample-conditional marginal distribution, which is expected to be uniform to distribute the total similarity across all data points evenly. On the other hand, $\mathbf{m}$, is the label-marginal distribution of the class assignments. Despite using a uniform distribution $\mathbf{m} = \mathbf{u}_K = \frac{1}{K}\mathbf{1}_K$ has provided satisfactory results on different computer vision tasks [3, 7], in our scenario, we constrain the solution to respect the observed label-marginal distribution on the calibration set, such that $\mathbf{m} = \frac{1}{N}\sum_1^N \mathbf{y}_i^{\text{ohe}}$, where $\mathbf{y}_i^{\text{ohe}}$ is the one-hot encoding of y_i.

Optimization. The objective in Eq. (5) is a linear program. However, its optimization is not straightforward, particularly regarding the computational complexity, exacerbated by increasing data points and categories. To alleviate this issue and provide a fast adaptation strategy, we resort to the Sinkhorn algorithm [10], which integrates an entropic constraint, enforcing a simple structure on the optimal regularized transport. Hence, the optimization problem becomes:

$$\max_{\mathbf{Q}\in\mathcal{Q}} tr(\mathbf{Q}^\top \mathbf{S}) + \varepsilon\mathcal{H}(\mathbf{Q}), \tag{7}$$

where $\mathcal{H}(\mathbf{Q})$ is the entropy, $\mathcal{H}(\mathbf{Q}) = -\sum_{ki} q_{ki}\log q_{ki}$, such that q are elements of $\mathbf{Q}$, and ε controls its weight. Now, the soft codes $\mathbf{Q}^*$ are the solution of the problem presented in Eq. (7) over the set $\mathcal{Q}$, which can be efficiently optimized with a few iterations as:

$$\mathbf{Q}^* = \text{Diag}(\mathbf{r}^{(t)})\mathbf{Q}^{(0)}\text{Diag}(\mathbf{c}^{(t)}). \tag{8}$$

The renormalization vectors are computed using a small number of matrix multiplications via the iterative Sinkhorn-Knopp algorithm, where in each iteration:

$$\mathbf{r}^{(t)} = \mathbf{m}/(\mathbf{Q}^{(0)}\mathbf{c}^{(t-1)}), \tag{9}$$

$$\mathbf{c}^{(t)} = \mathbf{u}_{(\mathbf{N+M})}/(\mathbf{Q}^{(0)}\mathbf{r}^{(t)}), \tag{10}$$

where $\mathbf{c}^{(0)} = \mathbf{1}_{(N+M)}$. Also, $\mathbf{Q}^{(0)}$ is initialized as $\mathbf{Q}^{(0)} = (\exp(\mathbf{S}/\tau)/\sum(\exp(\mathbf{S}/\tau))$, with τ representing a temperature scaling parameter that controls the strength of the entropic constraint in Eq. (7). Upon convergence, the matrix $\mathbf{Q}^*$ is normalized to produce soft class assignments that sum one for each sample, i.e., $\sum_k q_{ki}^* = \mathbf{1}_{(N+M)}{}^\top$.

Producing conformal sets through codes. Given the optimized matrix of codes, $\mathbf{Q}^*$, the final step consists of producing conformal sets from the obtained soft codes, $\boldsymbol{q}_i^*$ (columns in $\mathbf{Q}^*$), for each query sample. To do so, calibration, $\{(\boldsymbol{q}_i^{*\top}, y_i)\}_{i=1}^N$, and test, $\{(\boldsymbol{q}_i^{*\top})\}_{i=N+1}^{N+M}$, sets are separated again. Given an arbitrary non-conformity score, the predictive sets are created as detailed in Sec. 3.2: *i*) generating non-conformity scores from codes for calibration data, $s_i = \mathcal{S}(\boldsymbol{q}_i{}^\top, y_i)$, *ii*) finding the user-specified $1-\alpha$ quantile as in Eq. (3), and *iii*) creating conformal sets on test data based on such threshold following Eq. (4).

Algorithm 1 Conf-OT conformal prediction.

input: calibration dataset $\mathcal{D}_{\text{cal}} = \{(\boldsymbol{l}_i, y_i)\}_{i=1}^N$, query set $\mathcal{D}_{\text{test}} = \{(\boldsymbol{l}_i)\}_{i=N+1}^{N+M}$, non-conformity score function $\mathcal{S}$, error level α, entropic weight τ, iterations T.
// **Block 1.** - Transductive transfer learning.
// **Step 1.1.** - Init. optimal transport problem.
$\mathbf{S} \in \mathbb{R}^{K\times(N+M)} = [l_{ki}]_{k=1,i=1}^{k=K,i=N+M}$ // Sim. matrix.
$\mathbf{m} = \frac{1}{N}\sum_1^N \mathbf{y}_i^{\text{ohe}}$ // Label-marginal.
$\mathbf{u}_{(\mathbf{N+M})} = \frac{1}{(N+M)}\mathbf{1}_{(N+M)}$ // Sample marginal.
// **Step 1.2.** - Compute renormalization vectors.
$\mathbf{Q}^{(0)} = (\exp(\mathbf{S}/\tau)/\sum(\exp(\mathbf{S}/\tau))$ // Init. codes.
$\mathbf{c}^{(0)} = \mathbf{1}_{(N+M)}$ // Init. renormalization vector.
for t in $[1, \dots, T]$ **do**
  $\mathbf{r}^{(t)} = \mathbf{m}/(\mathbf{Q}^{(0)}\mathbf{c}^{(t-1)})$ // Eq. (9).
  $\mathbf{c}^{(t)} = \mathbf{u}_{(\mathbf{N+M})}/(\mathbf{Q}^{(0)}\mathbf{r}^{(t)})$ // Eq. (10).
end for
// **Step 1.3.** - Compute codes.
$\mathbf{Q}^* = \text{Diag}(\mathbf{r}^{(T)})\mathbf{Q}^{(0)}\text{Diag}(\mathbf{c}^{(T)})$ // Transport codes.
$\mathbf{Q}^* = \mathbf{Q}^*\text{Diag}(1/\sum_k q_{ki}^*)$ // Normalize.
// **Block 2.** - Conformal prediction.
$\mathcal{D}_{\text{cal}} = \{(\boldsymbol{q}_i^{*\top}, y_i)\}_{i=1}^N$, $\mathcal{D}_{\text{test}} = \{(\boldsymbol{q}_i^{*\top})\}_{i=N+1}^{N+M}$
// **Step 2.1.** - $1-\alpha$ non-conformity score quantile.
$\{s_i\}_{i=1}^N = \{\mathcal{S}(\boldsymbol{q}_i^{*\top}, y_i)\}_{i=1}^N$ // Non-conformity scores.
$\hat{s} \leftarrow \{s_i\}_{i=1}^N, \alpha$ // Search threshold - Eq. (3).
// **Step 2.2.** - Create query sets.
return: $\{\mathcal{C}(\boldsymbol{q}_i^{*\top})\}_{i=N+1}^M$ // Eq. (4).

Efficiency remarks. Immediate concerns might arise regarding the computational feasibility of Conf-OT. On the contrary, it is highly efficient, especially compared to other transductive pipelines. First, it operates over black-box logits, and second, it is *training-free*, requiring only a few iterations of the Sinkhorn algorithm with commodity resources. For the most extensive datasets, e.g., ImageNet ($K = 1,000$ and $(N+M) = 50,000$), the whole procedure requires only 1.1 seconds of additional overhead compared to its inductive counterpart (running on commodity CPUs). We extensively study its efficiency in Sec. 5.2 and Appendix G.6. Its robustness to different data ratios for calibration samples and query batch sizes is explored in Appendix G.7.

5. Experiments

5.1. Setup

Datasets. In this work, we leverage CLIP's zero-shot capabilities to deploy a large-scale benchmark of conformal inference strategies across a wide corpus of 15 datasets. Note that the main body of literature on conformal inference in vision [2, 51, 59, 69] is deployed on narrower scenarios using specialized models. In contrast, we use standard datasets for CLIP's zero- and few-shot adaptation, which gathers a heterogeneous number of tasks, from general objects to action recognition or fine-grained categories in specialized applications. These are: Imagenet [11], ImageNet-A [23], ImageNetV2 [50], ImageNet-R [24], ImageNet-Sketch [68], SUN397 [70], FGVCAircraft [37], EuroSAT [22], StanfordCars [27], Food101 [4], OxfordPets [47], Flowers102 [43], Caltech101 [16], DTD [8], and UCF101 [58]. We refer the reader to Appendix C for specific details on each task. The corresponding test partition from each dataset is employed for our conformal inference experiments by producing disjoint calibration and testing subsets.

Implementation details. We use CLIP [49] pre-trained models, using different backbones: ResNet-50 and ResNet-101 [21], and ViT-B/32, ViT-B/16, and ViT-L/14 [14]. Also, we experiment with MetaCLIP [71] ViT-B/16 and ViT-H/14 backbones. Unless otherwise indicated, ablation studies are performed with CLIP ViT-B/16. The text encoder from each model is used to produce class-wise prototypes for each downstream category by using standard templates and category names [18, 75], e.g., "`A photo of a [CLS]`.". Note that these templates are indicated in Appendix C. Then, logits for each sample are produced by computing the temperature-scaled cosine similarity as formalized in Sec. 3.1. The hyper-parameters of Conf-OT are fixed for all tasks: the entropic weight is set to $\tau = 1$, and the repetitions in the Sinkhorn algorithm are $T = 3$.

Baselines. Note that we find no clear candidate for tackling the proposed scenario: training-free black-box transductive adaptation of CLIP models over the logit space. Hence, we adjusted prior transductive approaches to operate in the logit space. First, TIM [5] is leveraged as a general transductive framework based on information maximization. Concretely, a modified version to incorporate the label-marginal distribution obtained from the calibration set using a Kullback-Leibler (KL) divergence, coined $\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}||\mathbf{m})}$, is employed, as well as a version using a uniform prior, $\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}||\mathbf{u}_K)}$. Second, we include the recently proposed TransCLIP [73], a GMM-based clustering method specially designed for VLMs. These baselines are formally introduced in Appendix F.

Conformal prediction algorithms. Three popular non-conformity scores for classification are assessed. In particular, we employ LAC [39], and two adaptive approaches, APS [51], and RAPS [2], to generate prediction sets at error rates of $\alpha \in \{0.1, 0.05\}$. We set the hyper-parameters in RAPS to $k_{\text{reg}} = 1$ and $\lambda = 0.001$. These values provided stable performance in [2]. Even though the authors in [2] provide different strategies for automatically fixing these values for set size or adaptiveness optimization, we avoid using additional validation splits for hyper-parameter tuning in our experimental setting.

Experimental protocol and metrics. The test subset from the target datasets is partitioned into equally-sized calibration and testing, following the standard split strategy in [2][3]. All experiments are repeated 20 times using different random seeds. We include discriminative performance metrics, such as Top-1 accuracy, and figures of merit typically employed in conformal prediction settings. Concretely, we compute the standard coverage ("Cov.") and average set size ("Size"), as well as class-conditioned coverage gap ("CCV"), which was recently proposed as a measure of adaptiveness [13]. These are formalized in Appendix D.

5.2. Main results

Enhancing SoTA conformal predictors. First, we compare the effect of Conf-OT with the base version of each non-conformity score using zero-shot predictions, i.e., no transfer learning. Results in Tab. 1 demonstrate the advantages of the proposed transductive approach to enhance recent conformal inference strategies. *Conf-OT provides consistent smaller set sizes for all conformal methods while maintaining the empirical coverage guarantees* for both $\alpha \in \{0.1, 0.05\}$. As a figure of its merit, set sizes consistently decrease in a relative ratio of nearly 20% compared to the base version. Also, class conditional coverage is consistently improved over 0.7 points when $\alpha = 0.1$ and 0.3 points when $\alpha = 0.05$. These results underscore the value of considering the structure of the unlabeled test samples during prediction to achieve better adaptability across many categories. Last, it is worth mentioning that the discriminative performance is enhanced notably by 2.6% for CLIP ResNet-50 and 2.9% for CLIP ViT-B/16. The positive performance of Conf-OT is also observed for additional CLIP and MetaCLIP encoders, whose results are provided in Appendix G.1. Results per dataset are in Appendix G.2.

Comparison to transductive baselines. Conf-OT is compared with relevant baselines in Tab. 2. The evaluation extends not only to the performance but also to the computational efficiency of such methods. The latter is of special relevance in the explored application since base conformal inference methods do not produce considerable overhead during inference and are designed to operate in real-world scenarios with limited hardware resources. The figures of merit in Tab. 2 indicate that *Conf-OT requires negligible additional inference times*. Also, *Conf-OT requires no spe-*

[3] Experiments using smaller data ratios are in Appendix G.7.

Method	$\alpha = 0.10$				$\alpha = 0.05$		
	Top-1↑	Cov	Size↓	CCV↓	Cov.	Size↓	CCV↓
CLIP ResNet-50							
LAC [39]	54.7	0.900	10.77	9.82	0.950	19.22	5.91
w/ Conf-OT	**57.3**$_{+2.6}$	0.900	**8.61**$_{-2.2}$	**9.15**$_{-0.7}$	0.951	**15.53**$_{-3.7}$	**5.61**$_{-0.3}$
APS [51]	54.7	0.900	16.35	8.36	0.950	26.50	5.34
w/ Conf-OT	**57.3**$_{+2.6}$	0.900	**12.94**$_{-3.4}$	**7.64**$_{-0.7}$	0.950	**20.96**$_{-5.5}$	**5.03**$_{-0.3}$
RAPS [2]	54.7	0.900	13.37	8.46	0.950	22.06	5.44
w/ Conf-OT	**57.3**$_{+2.6}$	0.900	**11.17**$_{-2.2}$	**7.72**$_{-0.7}$	0.950	**17.24**$_{-4.8}$	**5.19**$_{-0.3}$
CLIP ViT-B/16							
LAC [39]	63.8	0.899	5.52	10.37	0.950	10.24	6.14
w/ Conf-OT	**66.7**$_{+2.9}$	0.900	**4.40**$_{-1.1}$	**9.48**$_{-0.9}$	0.949	**7.99**$_{-2.3}$	**5.80**$_{-0.3}$
APS [51]	63.8	0.900	9.87	8.39	0.950	16.92	5.51
w/ Conf-OT	**66.7**$_{+2.9}$	0.899	**7.64**$_{-2.2}$	**7.44**$_{-1.0}$	0.949	**12.58**$_{-4.3}$	**5.09**$_{-0.4}$
RAPS [2]	63.8	0.900	8.12	8.50	0.950	12.66	5.52
w/ Conf-OT	**66.7**$_{+2.9}$	0.900	**6.68**$_{-1.4}$	**7.48**$_{-1.0}$	0.949	**10.11**$_{-2.6}$	**5.16**$_{-0.4}$

Table 1. **Conf-OT performance** atop popular non-conformity scores, i.e., LAC [39], APS [51], and RAPS [2]. Average performance across 15 datasets. "↓" indicates smaller values are better.

Method				$\alpha = 0.10$		
	Top-1↑	T	GPU	Cov.	Size↓	CCV↓
LAC [39]	63.8	**0.42**	-	0.899	5.52	10.37
$\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{u}_K)}$ [5]	64.7$_{+0.9}$	11.05	8.24	0.899	8.30$_{+2.8}$	10.41$_{+0.0}$
$\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{m})}$ [5]	65.0$_{+1.2}$	11.03	8.24	0.898	7.73$_{+2.2}$	10.89$_{+0.5}$
TransCLIP [73]	65.1$_{+1.3}$	12.00	12.2	**0.892**	5.76$_{+0.2}$	11.02$_{+0.7}$
Conf-OT	**66.7**$_{+2.9}$	0.61	-	0.900	**4.40**$_{-1.1}$	**9.48**$_{-0.9}$
APS [51]	63.8	**0.54**	-	0.900	9.87	8.39
$\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{u}_K)}$ [5]	64.7$_{+0.9}$	11.16	8.24	0.900	**7.24**$_{-2.6}$	9.32$_{+0.9}$
$\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{m})}$ [5]	65.0$_{+1.2}$	11.14	8.24	0.900	7.82$_{-2.1}$	9.38$_{+1.0}$
TransCLIP [73]	65.1$_{+1.3}$	12.12	12.2	**0.892**	8.27$_{-1.6}$	11.50$_{+3.1}$
Conf-OT	**66.7**$_{+2.9}$	0.72	-	0.899	7.64$_{-2.2}$	**7.44**$_{-1.0}$
RAPS [2]	63.8	**0.55**	-	0.900	8.12	8.50
$\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{u}_K)}$ [5]	64.7$_{+0.9}$	11.15	8.24	0.900	7.18$_{-0.9}$	9.32$_{+0.8}$
$\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}\|\|\mathbf{m})}$ [5]	65.0$_{+1.2}$	11.36	8.24	0.900	7.68$_{-0.4}$	9.42$_{+0.9}$
TransCLIP [73]	65.1$_{+1.3}$	12.12	12.3	0.899	7.17$_{-1.0}$	10.20$_{+1.7}$
Conf-OT	**66.7**$_{+2.9}$	0.74	-	0.900	**6.68**$_{-1.4}$	**7.48**$_{-1.0}$

Table 2. **Comparison to transductive baselines**. Results using CLIP ViT-B/16 on 15 datasets. "T" refers to runtime in seconds, and "GPU" to peak memory usage (Gb).

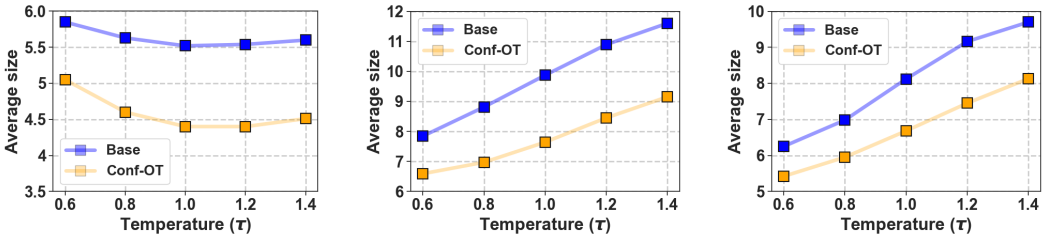
cific specialized hardware, as it can run on commodity resources. In contrast, popular transductive methods require considerable GPU modules and inference times. While being a much more efficient solution, Conf-OT also excels in performance. For instance, TIM and TransCLIP deteriorate the produced set sizes when using LAC conformal score. Regarding adaptive scores such as APS and RAPS, all methods provide improvements over the base version, with $\text{TIM}_{\text{KL}(\widehat{\mathbf{m}}||\mathbf{u}_K)}$ outperforming Conf-OT when using APS. We explain this positive performance of TIM+APS by the effect of the entropy minimization term on producing higher-confidence predictions, which positively affects APS [69]. However, such a positive trend is a mirage that does not hold when evaluated across additional backbones and coverage rates, as shown in Appendix G.3. Also, none of the considered methods improve the class-conditional coverage. Indeed, TransCLIP fails to provide the desired marginal coverage rate. Its Laplacian regularization term may cause this, as it is a neighborhood-based term that does not provide a joint optimization of calibration and test sets. The limitations of SoTA transductive methods enhance the qualities of the proposed solution. *Conf-OT is a training-free solution that provides consistently smaller conformal sets compared to SoTA, maintaining coverage guarantees*. As an additional bonus, Conf-OT also provides the best result regarding discrimination, i.e., Top-1 accuracy.

5.3. In-depth studies

In the following, we provide additional experiments to explore the conformal inference on VLMs in a more detailed manner, as well as key features of the proposed Conf-OT.

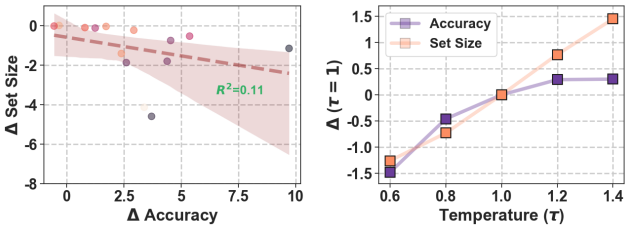
Complementary to temperature scaling. Conformal inference is usually related to other uncertainty frameworks, such as calibration. Notably, previous literature [2] tends to integrate calibration steps such as temperature scaling (TS). Recently, the authors in [69] explored the impact of temperature scaling on adaptive scores (APS, RAPS), observing that high-confidence predictions ($\tau < 1$) lead to smaller sets on average. The Sinkhorn optimal transport solver incorporates entropic constraints through a temperature-scaling parameter (see Algorithm 1), potentially affecting the conformal sets. Hence, we consider this aspect to deserve a specific study. Fig. 2 illustrates the joint effect of TS and Conf-OT. Results follow the observations in [69]: TS with $\tau < 1$ improves efficiency in adaptive methods, also when combined with Conf-OT. Notably, Conf-OT improvements are orthogonal to the ones produced by simply the sharpness of the probability distribution since it also controls other aspects, e.g., the label-marginal distribution in the overall assignment. Thus, while the effect of TS using non-adaptive scores such as LAC is absent, Conf-OT consistently improves. Based on these observations, we kept $\tau = 1$ for the entropic constraint weight in Conf-OT to provide a general framework for all non-conformity scores.

Is the improvement *only* on the discriminative aspect? One could argue that the better behavior of Conf-OT is explained by producing the largest probabilities on the correct class more often, i.e., discriminative performance, as observed in Top-1 accuracies improvement in Tab. 1. However, we have observed that this is not the case. For example, Fig. 3(a) shows a limited correlation between size and accuracy across datasets for LAC. Also, Fig. 3(b) illustrates an inverse trend when further optimizing the entropic constraint in adaptive conformal methods. Additionally, although all transductive baselines improve accuracy, they sometimes do not produce better conformal sets in Tab. 2. These observations indicate a disjoint behavior between discriminative and conformal inference figures.



(a) LAC [39] (b) APS [51] (c) RAPS [2]

Figure 2. **Entropic constraint (τ).** Conf-OT is compatible with recent observations [69] regarding the positive effect on set size of temperature scaling ($\tau < 1$) on adaptive scores (b,c). However, such behavior does not generalize to non-adaptive scores, i.e., LAC (a), whereas Conf-OT improves the performance atop all non-conformity scores. Results using CLIP ViT-B/16 on 15 datasets with $\alpha = 0.10$.



(a) (b)

Figure 3. **Accuracy *vs.* set size change (Δ) using Conf-OT.** (a) Correlation among datasets for LAC [39]. (b) Effect of the entropic constraint for RAPS [2]. Results using CLIP ViT-B/16 on 15 datasets with $\alpha = 0.10$. More information in Appendix G.4.

Method	Prior	$\alpha = 0.10$			
		Top-1↑	Cov.	Size↓	CCV↓
LAC [39]		63.8	0.899	5.52	10.37
w/ Conf-OT	$\mathbf{m} = \mathbf{u}_K$	65.5$_{+1.7}$	0.900	5.32$_{-0.2}$	9.87$_{-0.5}$
w/ Conf-OT	*Ours*	**66.7**$_{+2.9}$	0.900	**4.40**$_{-1.1}$	**9.48**$_{-0.9}$
APS [51]		63.8	0.900	9.87	8.39
w/ Conf-OT	$\mathbf{m} = \mathbf{u}_K$	65.5$_{+1.7}$	0.900	8.72$_{-1.2}$	**7.31**$_{-1.1}$
w/ Conf-OT	*Ours*	**66.7**$_{+2.9}$	0.899	**7.64**$_{-2.2}$	7.44$_{-1.0}$
RAPS [2]		63.8	0.900	8.12	8.50
w/ Conf-OT	$\mathbf{m} = \mathbf{u}_K$	65.5$_{+1.7}$	0.900	7.57$_{-0.6}$	**7.31**$_{-1.2}$
w/ Conf-OT	*Ours*	**66.7**$_{+2.9}$	0.900	**6.68**$_{-1.4}$	7.48$_{-1.0}$

Table 3. **Role of label-marginal prior** ($\mathbf{m}$ in Eq. (6)) in Conf-OT. Results using CLIP ViT-B/16 and averaged over 15 datasets.

Conf-OT components. The proposed approach presents a small number of tunable elements. First, as previously discussed, we fixed the entropic constraint weight to its standard value, $\tau = 1$. Second, we fixed the number of repetitions in the Sinkhorn algorithm to 3. These are enough for convergence, as shown in Appendix G.5. Last, Conf-OT uses the label-marginal distribution of the calibration set to constrain the optimal transport problem. Tab. 3 provides figures that showcase the importance of this element. It is worth mentioning that the potential of accessing this marginal distribution in transductive settings is not new. Indeed, oracle scenarios in image segmentation have also pointed in this direction [6]. Nevertheless, the standard conformal inference setting grants access to this information to ensure the assumption of data exchangeability [66]. The constraints of the Sinkhorn algorithm excel at efficiently incorporating such priors, especially compared to the other resource-demanding transduction baselines.

Data efficiency. We delve into this aspect of our transductive strategy in two measures: the calibration data ratio and robustness to small query sets. Specific numbers are in Appendix G.7. These demonstrate that *the efficiency of the sets produced by Conf-OT holds even in the most challenging scenarios*, e.g., using only 10% of data for calibration or receiving extremely small query sets of 8 or 16 images.

6. Limitations

In this work, we have explored the conformal prediction framework for zero-shot VLMs. To alleviate the absence of an adaptation stage, we have introduced a transductive setting to enhance the efficiency and adaptiveness of any conformal score by leveraging well-established knowledge in optimal transport. Our method is effective, but it presents some limitations. These are inherited from its transductive and conformal prediction nature. Particularly, it is valid under data exchangeability assumptions to guarantee the desired coverage, like any other conformal prediction method, and requires additional resources during inference, yet being $\times 15$ faster than other transductive approaches.

Acknowledgments

This work was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC). We also thank Calcul Québec and Compute Canada.

References

[1] Vladimir Vapnik Alex Gammerman, Volodya Vovk. Learning by transduction. In *Conference on Uncertainty in Artificial Intelligence*, pages 148–156, 1998. 1, 3, 12

[2] Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 3, 4, 6, 7, 8, 12, 13, 15, 16, 18, 19, 20, 21, 22, 23, 24, 25, 26

[3] YM Asano, C Rupprecht, and A Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *International Conference on Learning Representations (ICLR)*, 2020. 5

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, 2014. 6, 13

[5] Malik Boudiaf, Ziko Imtiaz Masud, Jérôme Rony, Jose Dolz, Pablo Piantanida, and Ismail Ben Ayed. Transductive information maximization for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1–13, 2020. 3, 6, 7, 14, 15, 19

[6] Malik Boudiaf, Hoel Kervadec, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13979–13988, 2021. 8

[7] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision (ECCV)*, pages 132–149, 2018. 5, 16

[8] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3606–3613, 2014. 6, 13

[9] Alvaro H.C. Correia, Fabio Valerio Massoli, Christos Louizos, and Arash Behboodi. An information theoretic perspective on conformal prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3, 12

[10] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013. 4, 5

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 1, 6, 13

[12] Guneet Singh Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. In *International Conference on Learning Representations (ICLR)*, 2020. 3

[13] Tiffany Ding, Anastasios Angelopoulos, Stephen Bates, Michael Jordan, and Ryan J Tibshirani. Class-conditional conformal prediction with many classes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 4, 6, 12, 13

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*, 2021. 6

[15] Bat-Sheva Einbinder, Yaniv Romano, Matteo Sesia, and Yanfei Zhou. Training uncertainty-aware classifiers with conformalized deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3, 12

[16] Li 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Worskshops (CVPRW)*, pages 178–178, 2004. 6, 13

[17] Mengdi Gao, Hongyang Jiang, Lei Zhu, Zhe Jiang, Mufeng Geng, Qiushi Ren, and Yanye Lu. Discriminative ensemble meta-learning with co-regularization for rare fundus diseases diagnosis. *Medical Image Analysis*, 89:102884, 2023. 2, 12

[18] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision (IJCV)*, 2023. 1, 2, 3, 4, 6, 13

[19] Ulysse Gazin, Gilles Blanchard, and Etienne Roquain. Transductive conformal inference with adaptive scores. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 348–360, 2024. 3, 12

[20] Changsheng Xu Hantao Yao, Rui Zhang. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[22] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pages 3606–3613, 2018. 6, 13

[23] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15262–15271, 2019. 6, 13

[24] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, page 8340–8349, 2021. 6, 13

[25] Jianguo Huang, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. Conformal prediction for deep classifier via label ranking. In *International Conference on Machine Learning (ICML)*, pages 20331–20347, 2024. 1, 4

[26] Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. Lp++: A surprisingly strong linear probe for few-shot clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23773–23782, 2024. 1, 2

[27] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 3498–3505, 2012. 6, 13

[28] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Preprint*, 2009. 1

[29] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018. 4

[30] Xiaomeng Li, Lequan Yu, Yueming Jin, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Difficulty-aware meta-learning for rare disease diagnosis. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 357–366, 2020. 2, 12

[31] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7061–7070, 2023. 1

[32] Xiwen Liang, Yangxin Wu, Jianhua Han, Hang Xu, Chunjing Xu, and Xiaodan Liang. Effective adaptation in multi-task co-training for unified autonomous driving. *Advances in Neural Information Processing Systems (NeurIPS)*, 35: 19645–19658, 2022. 1

[33] Chuang Lin, Peize Sun, Yi Jiang, Ping Luo, Lizhen Qu, Gholamreza Haffari, Zehuan Yuan, and Jianfei Cai. Learning object-language alignments for open-vocabulary object detection. In *International Conference on Learning Representations (ICML)*, 2023. 1

[34] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2

[35] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21152–21164, 2023. 1

[36] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning (ICML)*, pages 23033–23044. PMLR, 2023. 1

[37] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. 2013. 6, 13

[38] Segolene Martin, Yunshi Huang, Fereshteh Shakeri, Jean-Christophe Pesquet, and Ismail Ben Ayed. Transductive zero-shot and few-shot clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28816–28826, 2024. 3

[39] Jing Lei Mauricio Sadinle and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525): 223–234, 2019. 1, 3, 4, 6, 7, 8, 12, 15, 18, 19, 20, 21, 22, 23, 24, 25, 26

[40] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *International Conference on Learning Representations (ICLR)*, pages 1–17, 2023. 2

[41] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024. 1

[42] Balamurali Murugesan, Julio Silva-Rodriguez, Ismail Ben Ayed, and Jose Dolz. Robust calibration of large vision-language adapters. In *European Conference on Computer Vision (ECCV)*, pages 1–19, 2021. 1, 2

[43] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 6, 13

[44] Zachary Novack, Julian McAuley, Zachary Chase Lipton, and Saurabh Garg. Chils: Zero-shot image classification with hierarchical label sets. In *International Conference on Machine Learning (ICML)*, pages 26342–26362, 2023. 1

[45] Changdae Oh, Mijoo Kim, Hyesu Lim, Junhyeok Park, Euiseog Jeong, Zhi-Qi Cheng, and Kyungwoo Song. Towards calibrated robust fine-tuning of vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 1

[46] Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *European Conference on Machine Learning (ECML)*, pages 345–356, 2002. 1, 3, 4, 12

[47] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 3498–3505, 2012. 6, 13

[48] Kostas Proedrou, Ilia Nouretdinov, Volodya Vovk, and Alex Gammerman. Transductive confidence machines for pattern recognition. In *European Conference on Machine Learning (ECML)*, pages 381–390, 2002. 1, 3, 12

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 1, 2, 3, 6, 13

[50] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning (ICML)*, pages 5389–5400, 2019. 6, 13

[51] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3581–3591, 2020. 1, 3, 4, 6, 7, 8, 12, 15, 16, 18, 19, 20, 21, 22, 23, 24, 25, 26

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 1

[53] Craig Saunders, Alexander Gammerman, and Volodya Vovk. Transduction with confidence and credibility. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 722–726, 1999. 1, 3, 12

[54] Fereshteh Shakeri, Yunshi Huang, Julio Silva-Rodríguez, Houda Bahig, An Tang, Jose Dolz, and Ismail Ben Ayed. Few-shot adaptation of medical vision-language models. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 553–563, 2024. 1, 2, 12

[55] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:14274–14289, 2022. 1, 2

[56] Julio Silva-Rodríguez, Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. A closer look at the few-shot adaptation of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23681–23690, 2024. 1, 2, 4, 13

[57] Julio Silva-Rodríguez, Hadi Chakor, Riadh Kobbi, Jose Dolz, and Ismail Ben Ayed. A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. *Medical Image Analysis*, 99:103357, 2025. 1

[58] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. 2012. 6, 13

[59] David Stutz, Krishnamurthy Dj Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers. In *International Conference on Learning Representations (ICLR)*, 2022. 1, 3, 4, 6, 12, 13

[60] Weijie Tu, Weijian Deng, Dylan Campbell, Stephen Gould, and Tom Gedeon. An empirical study into what matters for calibrating vision-language models. In *International Conference on Machine Learning (ICML)*, 2024. 1

[61] Vishaal Udandarao, Ameya Prabhu, Adhiraj Ghosh, Yash Sharma, Philip H. S. Torr, Adel Bibi, Samuel Albanie, and Matthias Bethge. No "zero-shot" without exponential data: Pretraining concept frequency determines multimodal model performance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[62] Olivier Veilleux, Malik Boudiaf, Pablo Piantanida, and Ismail Ben Ayed. Realistic evaluation of transductive few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 9290–9302, 2021. 3

[63] Cédric Villani. *Optimal transport: old and new*. Springer, 2009. 4

[64] Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning*, pages 475–490, 2012. 3, 12

[65] Vladimir Vovk. Transductive conformal predictors. In *Artificial Intelligence Applications and Innovations*, pages 348–360, 2013. 3

[66] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005. 1, 2, 3, 4, 8, 12

[67] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[68] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 6, 13

[69] Huajun Xi, Jianguo Huang, Kangdao Liu, Lei Feng, and Hongxin Wei. Does confidence calibration improve conformal prediction? In *arXiv preprint arXiv:2402.04344*, 2024. 3, 4, 6, 7, 8, 13, 16

[70] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3485–3492, 2010. 6, 13

[71] Hu Xu, Saining Xie, Xiaoqing Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying CLIP data. In *International Conference on Learning Representations (ICLR)*, 2024. 6

[72] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10899–10909, 2023. 1

[73] Maxime Zanella, Benoît Gérin, and Ismail Ben Ayed. Boosting vision-language models with transduction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3, 6, 7, 14, 15, 19

[74] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. In *European Conference on Computer Vision (ECCV)*, pages 1–19, 2022. 1, 2

[75] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022. 1, 2, 3, 6, 13

[76] Hao Zhu and Piotr Koniusz. Ease: Unsupervised discriminant subspace learning for transductive few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9068–9078, 2022. 3