Scaling Full Conformal Image Classifiers

Julio Silva-Rodríguez
Computer Vision Lab, ETH Zürich
jusilva@ethz.ch

Ender Konukoglu
Computer Vision Lab, ETH Zürich

Abstract

Conformal prediction provides set-valued predictions with distribution-free coverage guarantees, making it attractive for high-stakes image classification. However, split conformal prediction is data-inefficient, while full conformal prediction (FCP), despite its stronger statistical efficiency, is computationally prohibitive at scale because it requires candidate-specific model refits at test time. We address this limitation by leveraging zero-shot vision-language models (VLMs) to guide scalable FCP in large label spaces. We introduce Targeted Full Conformal Prediction (T-FCP), which uses a lightweight inductive conformal predictor to prune unlikely labels and applies FCP only to the remaining candidates, reducing computation while retaining the formal guarantee of the combined conformal procedure. We further propose Stabilized Online LDA (SO-LDA), an efficient VLM adaptation solver based on rank-one inverse-covariance updates. Across multiple benchmarks, including ImageNet, T-FCP enables practical full-conformal image classification with modest test-time overhead, yielding efficient prediction sets and more stable empirical coverage than split conformal alternatives. Code is available: T-FCP

1 Introduction

In image classification, an input image $\mathbf{x} \in \mathcal{X}$ is mapped to a label space $\mathcal{Y} = \{1, \dots, C\}$ via a trained model $\pi_\theta : \mathcal{X} \to \Delta_C$, where Δ_C denotes the probability simplex, and θ the parameters. Predictions are typically turned into decisions by selecting the label with the highest score. However, such a procedure does not provide information about model certainty, so potential errors or uncertain cases might remain unnoticed. While modern neural networks are increasingly adopted in high-stakes computer vision applications, e.g., medical imaging [32] or autonomous driving [30], there is a growing need to provide users with trustworthy operational decisions with appropriate guarantees.

Conformal prediction (CP) [54, 1, 43] is a machine learning framework that creates set-valued predictions with formal, distribution-free coverage guarantees. In the classification setting, CP constructs a prediction set $\mathcal{C}(\mathbf{x}) \subseteq \mathcal{Y}$ through a set-valued mapping $\mathcal{C} : \mathcal{X} \to 2^{\mathcal{Y}}$, such that the following marginal coverage property holds:

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

where $\alpha \in (0, 1)$ is a user-specified error rate. This guarantee ensures, under mild exchangeability assumptions, that the true label lies within the prediction set with probability at least $1 - \alpha$ under the target data distribution. Importantly, the size of $\mathcal{C}(\mathbf{x})$ adapts to the uncertainty of the prediction, such that inputs more ambiguous according to the model typically yield larger sets

The construction of $\mathcal{C}$ is data-driven, as it relies on a *labeled calibration dataset*, $\mathcal{D}_N = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$. Specifically, CP requires defining a nonconformity score function $\mathcal{S} : (\mathcal{X} \times \mathcal{Y}) \times \theta_{\mathcal{D}_*} \to \mathbb{R}$, where $\mathcal{S}((\mathbf{x}, y); \theta_{\mathcal{D}_*})$ quantifies how atypical a candidate label y is for a given input $\mathbf{x}$ relative to a model trained on a reference dataset $\mathcal{D}_*$. CP uses the calibration data to determine a rejection threshold via an

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

empirical quantile search on the observed score distribution $s_i = \mathcal{S}((\mathbf{x}_i, y_i); \theta_{\mathcal{D}_*})$ for $i = 1, \ldots, N$:

$$\hat{q}_\alpha = \text{Quantile}(\mathcal{D}_N; \mathcal{S}; \alpha) = \inf \left\{ s \in \mathbb{R} : \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{s_i \leq s\} \geq \frac{\lceil (N+1)(1-\alpha) \rceil}{N} \right\}, \quad (2)$$

which corresponds to a finite-sample corrected $(1-\alpha)$-quantile of the calibration nonconformity scores, and is used for a new test input, $\mathbf{x}_{N+1}$, for constructing the prediction set:

$$\mathcal{C}(\mathbf{x}_{N+1}) = \{y \in \mathcal{Y} : \mathcal{S}((\mathbf{x}_{N+1}, y); \theta_{\mathcal{D}_*}) \leq \hat{q}_\alpha\}. \quad (3)$$

By definition, this set contains all labels whose nonconformity scores are sufficiently small, and therefore *conform* to the calibration distribution.

Designing valid scoring functions. Several choices of $\mathcal{S}$ exist in the literature, which in turn influence the structure of the resulting prediction sets, typically building upon predicted class probabilities. As an illustrative example, one may define $\mathcal{S}(\mathbf{x}, y; \theta) = 1 - \pi_\theta(\mathbf{x})_y$, which assigns small scores to labels with high predicted probability and larger scores to less likely alternatives. A key formal requirement for conformal validity, however, is that the resulting nonconformity scores are *exchangeable* when evaluated on the calibration and test samples. Formally, consider the augmented dataset, $\mathcal{D}_{N+1} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N+1}$, assumed to be exchangeable. Let $s_i = \mathcal{S}(\mathbf{x}_i, y_i; \theta)$ denote the corresponding scores. Then, for any permutation σ of $\{1, \ldots, N+1\}$, it holds that

$$(s_1, \ldots, s_{N+1}) \overset{d}{=} (s_{\sigma(1)}, \ldots, s_{\sigma(N+1)}), \quad (4)$$

i.e., the joint distribution of the scores is *permutation-invariant*. This condition ensures that the test score is statistically indistinguishable from calibration scores, yielding the marginal coverage guarantee in Eq. (1). In practice, *Inductive Conformal Prediction* (ICP) [37, 54] enforces this requirement by learning the model parameters θ on data that is disjoint from the calibration set, thereby ensuring that the scores $\{s_i\}_{i=1}^{N+1}$ are computed using a fixed, calibration data-independent model. When the calibration set is obtained by splitting the available labeled data, this procedure is commonly referred to as *Split Conformal Prediction* (SCP) [37, 54, 44, 53].

The data-efficiency challenge. Despite the SCP being the most common procedure for image classification, it is also inherently data-inefficient, as only a fraction of the available data contributes to training and to the empirical quantile estimation. This limitation is nontrivial: the marginal coverage guarantee holds only in expectation over the calibration samples, introducing finite-sample variability (i.e., a beta-binomial distribution on α and N [53, 21, 34]) that may lead to noticeable deviations from the target coverage when N is small. For instance, *a user specifying* $\alpha = 0.1$ *may observe empirical coverage closer to* 85% *instead of* 90% *in practice.*

Full Conformal Prediction (FCP) [43] is, in contrast, a more appealing alternative in terms of data efficiency, as it leverages the entire available calibration labeled data for both model fitting and quantile estimation without splits. The key idea is to maintain the score exchangeability requirement in Eq. 4 by ensuring that the fitting function is symmetric with respect to the extended dataset, $\mathcal{D}_{N+1}$, which means that for all $(\mathbf{x}_i, y_i) \in \mathcal{D}_{N+1}$, and any permutation σ:

$$\mathcal{S}((\mathbf{x}_i, y_i); \theta_{\mathcal{D}_{N+1}}) = \mathcal{S}((\mathbf{x}_i, y_i); \theta_{\sigma(\mathcal{D}_{N+1})}). \quad (5)$$

Since y_{N+1} is unknown, the FCP procedure requires testing the conformity of each new test sample $\mathbf{x}$ and candidate label $y \in \mathcal{Y}$. Each augmented dataset $\mathcal{D}^y_{N+1} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N), (\mathbf{x}_{N+1}, y)\}$ is used to fit the parameters of the scoring function, $\theta_{\mathcal{D}^y_{N+1}}$, which produces its respective nonconformity scores, $s^y_i = \mathcal{S}((\mathbf{x}_i, y_i); \theta_{\mathcal{D}^y_{N+1}})$. The (1-$\alpha$)-quantile using such score function is obtained from the same calibration set, $\hat{q}^y_\alpha$, and is used as a rejection criterion to construct valid conformal sets:

$$\mathcal{C}_{\text{FCP}}(\mathbf{x}_{N+1}) = \{y \in \mathcal{Y} : \mathcal{S}((\mathbf{x}_{N+1}, y); \theta_{\mathcal{D}^y_{N+1}}) \leq \hat{q}^y_\alpha\}. \quad (6)$$

Despite being more data-efficient, the FCP procedure creates two computational bottlenecks at test time: (**B1**) testing each potential label for an incoming test point, and (**B2**) the cost of training each classifier during testing. Consequently, especially considering standard large-scale computer vision benchmarks, e.g., ImageNet, where $C = 1000$, FCP has been systematically neglected in favor of SCP, at the cost of assuming larger annotation efforts or wider coverage distributions.

Contributions. In this work, we aim to enable the construction of full conformal image classifiers at scale. Our main observation is that in a large label space, despite not being intrinsically ordered,

many classes may be weakly related to each other, e.g., "husky" and "toaster", while some groups of categories might be more similar, e.g., "husky", "malamute", "Samoyed", or "Greater Swiss Mountain dog" (*real examples from ImageNet*). Therefore, if an initial proxy were available to guide the FCP procedure toward the most likely labels for a new test image, the total runtime could be significantly reduced. Hence, we propose exploiting the open-vocabulary capabilities of pre-trained vision-language models (VLMs), such as CLIP [39]. These provide zero-shot class prototypes via text descriptions for the target label domain, which, despite not being specialized in the downstream task, serve as proxies in the model's embedding space to discard non-similar categories that are not considered in the FCP procedure for each test image. Technically, our contributions are:

- We formalize *Targeted Full Conformal Prediction* (T-FCP), a procedure that only requires running FCP on a subset of the label space (**B1**), defined by an ICP procedure guided by zero-shot class prototypes with provable coverage guarantees without requiring splits on calibration data.
- To address the cost of multiple model fits (**B2**), we introduce SO-LDA, an online Linear Discriminant Analysis solver for VLM adaptation that performs stable incremental class-prototype updates together with rank-one inverse-covariance updates. SO-LDA achieves performance close to expensive gradient-based solvers while being orders of magnitude faster.

Extensive experiments demonstrate that our framework enables full conformal procedures on large-scale datasets with a modest latency overhead (e.g., ~ 30 ms/image for ImageNet), yielding efficient sets and more stable coverage distributions than split conformal procedures.

2 Related Work

Full conformal prediction. Recent work on full conformal prediction has focused on reducing its computational burden, e.g., by limiting the candidate label space evaluated at test time. In regression, this has been achieved by exploiting the natural ordering of the response space, either through discretization [28, 27] or homotopy/continuation methods that track the solution path of the fitted model as the candidate response varies [27, 26, 17, 29]. Classification problems, however, do not directly admit these strategies. Existing work for classification has instead focused on online classifiers with decoupled test-time updates, including exact permutation-invariant updates for kNN and SVMs [6], as well as inexact updates for gradient-based solvers via influence functions [35].

Conformal prediction for image classification. Most neural-network-based image classification works follow SCP [11, 51, 2, 9, 20, 5, 10], improving either training objectives for set efficiency [11, 51] or nonconformity scores for adaptiveness, class-conditional coverage, and robustness [2, 5, 9, 10]. In contrast, calibration-data efficiency and empirical coverage variability have received less attention, as these methods typically rely on in-domain training data and separate calibration splits.

Conformal prediction for zero-shot models. Recent works have explored split conformal prediction for vision foundation models [13] and, more specifically, vision-language models (VLMs) [47]. Several methods improve VLM adaptation within conformal procedures using unsupervised transductive adaptation [47, 46, 4]. Closer to our work, [49] explored full conformal classifiers using closed-form linear solvers in the VLM embedding space. However, [49] focused on small-scale datasets ($C < 50$). As we show later, their proposed kNN-style solver, SS-Text, does not scale well to large label spaces and fine-grained classification tasks.

Combining conformal predictors. Related to our work are methods that combine or cascade multiple split conformal procedures [14, 52, 42]. For example, conformal cascades progressively prune candidate sets obtained with different nonconformity scores [14]. Other works derive combined coverage guarantees, either through multiplicative bounds under independence assumptions [52], or by allocating individual error rates using common corrections such as Bonferroni [52, 42]. These approaches have been used to achieve goals such as label-dependent guarantees for object detection [52] and joint guarantees across multiple outputs [42].

3 Methods

3.1 Background on zero-shot image classifiers

Let $f_\theta : \mathcal{X} \to \mathbb{R}^F$ denote a visual encoder that maps an input image $\mathbf{x}$ to a feature vector $\mathbf{v} = f_\theta(\mathbf{x}) \in \mathbb{R}^F$. We consider contrastive vision-language models (VLMs), following CLIP [39], as our

zero-shot models. These provide class prototypes from class descriptions (e.g., "an image of a `[class name]`"), via its text encoder g_ϕ, producing textual embeddings $\mathbf{t}_y = g_\phi(\text{prompt}_y) \in \mathbb{R}^F$. These embeddings can be directly used as class prototypes, i.e., $\mathbf{W}^0 = (\mathbf{t}_y)_{y=1}^C$, by measuring the similarity between ℓ_2-normalized image embeddings and class prototypes, yielding class probabilities:

$$\pi_{\mathbf{W}^0}(\mathbf{x}_i) = (p_{i,y}^0)_{y=1}^C \in \Delta_C, \quad p_{i,y}^0 = \frac{\exp\left(\mathbf{v}_i^\top \mathbf{w}_y^0/\tau\right)}{\sum_{j\in\mathcal{Y}} \exp\left(\mathbf{v}_i^\top \mathbf{w}_j^0/\tau\right)}, \tag{7}$$

3.2 Targeted Full Conformal Prediction

First, we aim to leverage zero-shot models to address the first bottleneck in full conformal prediction procedures (**B1**): the need to iterate over the entire label space for each incoming test sample.

Combining conformal procedures. Our approach is based on the observation that multiple valid conformal predictors can be combined while maintaining statistical guarantees. Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be two conformal predictors constructed using error levels α_1 and α_2, respectively. That is, for a test point $(\mathbf{x}, Y)$, and under the standard exchangeability assumptions:

$$\mathbb{P}(Y \in \mathcal{C}_1(\mathbf{x})) \geq 1 - \alpha_1, \quad \mathbb{P}(Y \in \mathcal{C}_2(\mathbf{x})) \geq 1 - \alpha_2. \tag{8}$$

We define a combined prediction set as the intersection

$$\mathcal{C}_{1\cap 2}(\mathbf{x}) = \mathcal{C}_1(\mathbf{x}) \cap \mathcal{C}_2(\mathbf{x}). \tag{9}$$

Proposition 1. *The prediction set $\mathcal{C}_{1\cap 2}(\mathbf{x})$ satisfies*

$$\mathbb{P}\big(Y \in \mathcal{C}_{1\cap 2}(\mathbf{x})\big) \geq 1 - (\alpha_1 + \alpha_2). \tag{10}$$

The proof follows directly from De Morgan's law and the union bound, and is given in Appendix B.

Remark 1 (Tightness and conservativeness)**.** *The bound in Proposition 1 is a worst-case guarantee and may be conservative in practice. It becomes tight when the two miscoverage events are disjoint and have probabilities α_1 and α_2. When the error events overlap, the intersection can have coverage higher than $1 - (\alpha_1 + \alpha_2)$. If the individual miscoverage probabilities are exactly α_1 and α_2, the coverage of the intersection lies between $1 - (\alpha_1 + \alpha_2)$ and $1 - \max\{\alpha_1, \alpha_2\}$.*

Targeted conformal prediction via zero-shot models. Let us now introduce a labeled calibration set consisting of embedding representations, $\mathcal{D}_N = \{(\mathbf{v}_i, y_i)\}_{i=1}^N$. A set of zero-shot linear weights, $\mathbf{W}^0$, is presented, and the goal is to create conformal sets with coverage guarantees from new incoming test images, using their embedding, $\mathbf{v}_{N+1}$. Let us now define an ICP procedure, $\mathcal{C}_{\text{ICP}}$, whose nonconformity score function leverages the zero-shot weights, $\mathcal{S}(\mathbf{x}, y; \mathbf{W}^0)$, and produces conformal sets with an error rate of α_{ICP} using the whole calibration set. Also, let us denote an FCP procedure, $\mathcal{C}_{\text{FCP}}$, that uses a more expensive scoring function by fitting the classifier on the extended dataset $\mathcal{S}((\mathbf{x}, y); \mathbf{W}^{\mathcal{D}_{N+1}^y})$, operating at an error level α_{FCP}. We define the *Targeted Full Conformal Prediction* (T-FCP) sets as the intersection of both procedures, i.e., $\mathcal{C}_{\text{T-FCP}} = \mathcal{C}_{\text{ICP}\cap\text{FCP}}$, such that, according to Proposition 1, the following marginal coverage property holds:

$$\mathbb{P}\big(Y \in \mathcal{C}_{\text{T-FCP}}(\mathbf{x})\big) \geq 1 - (\alpha_{ICP} + \alpha_{FCP}). \tag{11}$$

By construction, any label excluded by $\mathcal{C}_{\text{ICP}}$ for a test sample cannot appear in $\mathcal{C}_{\text{T-FCP}}$. Therefore, we can use $\mathcal{C}_{\text{ICP}}$ as a *pruning stage* of highly unlikely labels to be included in the conformal set, and only run the expensive FCP procedure, $\mathcal{C}_{\text{FCP}}$, on the surviving, i.e., *target labels*, while maintaining theoretical guarantees. Specifically, for a given target error rate α, we can set $\alpha_{FCP} = \alpha - \alpha_{ICP}$. The parameter α_{ICP} controls a computational-statistical tradeoff: larger values typically yield smaller ICP candidate sets and therefore lower FCP cost, but allocate less error budget to the full conformal stage. In our experiments, we find that $\alpha_{\text{ICP}} = 0.5\%$ already provides substantial computational efficiency gains while keeping the additional conservativeness limited in practice.

3.3 SO-LDA: Stabilized Online LDA

Second, we exploit the capabilities of zero-shot models to alleviate the second bottleneck in full conformal prediction (**B2**): the computational cost of fitting the classifier for each extended dataset.

Transfer learning. Instead of training a deep network for each extended dataset, we leverage the rich embedding representation of pre-trained vision-language models, similarly to [49]. Given a dataset with labeled visual examples, and its corresponding embeddings, $\mathcal{D}_N = \{(\mathbf{v}_i, y_i)\}_{i=1}^N$, a common strategy is to learn task-specific linear class prototypes $\mathbf{W}^{\mathcal{D}_N}$, that replace its zero-shot counterpart in Eq. 7 to produce softmax class scores.

Online LDA. While several linear probe solvers have been proposed in the literature [16, 45, 56], gradient-descent-based minimization of softmax cross-entropy is generally the best-performing solution [31, 45]. However, its computational cost ($\mathcal{O}(TNFC)$, with T the number of iterations) makes it infeasible for full conformal prediction. We therefore turn our attention to training-free solvers [59, 56, 49, 48], particularly those that enable decoupled updates between calibration samples and new incoming test points. Specifically, we build upon the Linear Discriminant Analysis (LDA) solver [15], proposed in [56] for VLMs, which models class-conditional feature distributions as Gaussian with a shared covariance matrix. In its simplified form, by dropping the bias term for consistency with Eq.7, the classifier weights are given by:

$$\mathbf{w}_y = \mathbf{\Sigma}^{-1}\mu_y \tag{12}$$

where μ_y are the class centers, and Σ^{-1} the inverse covariance matrix.

Given a labeled set $\mathcal{D}_N$, and let $N_y = \sum_{i\in\mathcal{D}_N} \mathbf{1}\{y_i = y\}$ denote the number of samples for each category, then the model parameters can be estimated as follows:

$$\mu_{y(N)} = \frac{1}{N_y}\sum_{i\in\mathcal{D}_N} \mathbf{1}\{y_i = y\}\mathbf{v}_i, \quad \mathbf{\Sigma}_{(N)}^{-1} = \mathbf{S}_{(N)}^{-1} = (\frac{1}{N}\sum_{i\in\mathcal{D}_N} \mathbf{z}_i\mathbf{z}_i^\top)^{-1}, \tag{13}$$

where $\mathbf{z}_i = \mathbf{v}_i - \mu_{y=y_i(N)}$ are the class-centered features. To avoid numerical instabilities, we use diagonal-loading stabilization, $\mathbf{S}_{(N)}^{\text{reg}} = \mathbf{S}_{(N)} + \lambda_{\text{REG}}\,\text{Diag}(\mathbf{S}_{(N)})$, following the common use of shrinkage for high-dimensional precision estimation [24, 56]. The loading term is estimated once from the calibration covariance and kept fixed during candidate augmentations, preserving efficient rank-one updates. Appendix C compares it with fixed-ridge and full recomputation variants.

Given a new incoming sample, $(\mathbf{v}_{N+1}, y_{N+1})$, the model parameters can be updated [25], instead of being recalculated again. Particularly, for the covariance matrix updates, we can consider the Sherman–Morrison formula [18] to compute the inverse of its rank-one update efficiently:

$$\mu_{y(N+1)} = \frac{N_y \cdot \mu_{y(N)} + \mathbf{v}_{N+1}\mathbf{1}\{y_{N+1} = y\}}{N_y + \mathbf{1}\{y_{N+1} = y\}}, \tag{14}$$

$$\mathbf{\Sigma}_{(N+1)}^{-1} = \frac{N+1}{N}\left(\mathbf{\Sigma}_{(N)}^{-1} - \frac{\mathbf{\Sigma}_{(N)}^{-1}\mathbf{z}_{N+1}\mathbf{z}_{N+1}^\top\mathbf{\Sigma}_{(N)}^{-1}}{N + \mathbf{z}_{N+1}^\top\mathbf{\Sigma}_{(N)}^{-1}\mathbf{z}_{N+1}}\right). \tag{15}$$

In full conformal procedures, the parameters are first estimated from the calibration data and then updated for each augmented dataset $\mathcal{D}_{N+1}^y$. This reduces the per-label update cost to $\mathcal{O}(F^2)$, avoiding repeated matrix inversion ($\mathcal{O}(F^3)$) and enabling scalable conformal inference compared to recomputing the LDA solution from scratch ($\mathcal{O}(NF^2 + F^3)$).

Enabling stable online updates through zero-shot prototypes. A practical difficulty of online LDA in the full conformal setting is that the residual vectors $\mathbf{z}_i$ for samples in $\mathcal{D}_N$ are computed using statistics estimated before the candidate test sample is observed, whereas the test sample itself is evaluated under updated statistics. This creates an asymmetry between the approximate online update (Eq. 15) and a full refit on the augmented dataset. To mitigate this issue, we introduce **S**tabilized **O**nline **LDA** (**SO-LDA**), including the following adjustments:

$$\mathbf{z}_i = \mathbf{v}_i - \mathbf{W}_{y_i}^{(0)}, \tag{16}$$

$$\mu_y^{\text{SO-LDA}} = \mu_y^{(N+1)} + \lambda_{\text{TEXT}}\mathbf{W}_y^{(0)}. \tag{17}$$

According to Eq. 16, the feature-centering anchors used for covariance estimation no longer depend on the calibration samples, but on the zero-shot prototypes. Therefore, the residual centering rule remains invariant under candidate augmentation and avoids the asymmetry induced by recomputing visual class centers differently for calibration and test samples. Additionally, following standard practices in VLM adaptation [45, 48, 49], Eq. 17 introduces a text-guided prior (weighted by λ_{TEXT}) on the updated class means by biasing them toward the textual prototypes before ℓ_2-normalization and score prediction in Eq. 12. This improves stability, especially in low-data regimes.

4 Experiments

4.1 Setup

Datasets. We evaluate the proposed procedure in the standard 11 datasets used for CLIP zero- and few-shot transfer evaluation [45, 56]. These are: ImageNet [8], SUN397 [57], FGVCAircraft [33], EuroSAT [19], StanfordCars [23], Food101 [3], OxfordPets [38], Flowers102 [36], Caltech101 [12], DTD [7], and UCF101 [50]. These gather a heterogeneous set of general, fine-grained, and specialized domain classification tasks, several of which have hundreds of categories. We refer to Appendix D for specific details on the number of categories and tasks.

Calibration sets. The corresponding test partition from each dataset is used in our conformal experiments to produce disjoint calibration and test subsets. Specifically, samples of size $N = C \times K$, with K the so-called number of shots in the transfer learning literature [60, 16, 45], are retrieved for calibration. By default, we set $K = 16$, a standard data regime in transfer learning from VLMs [60, 16, 45], and present ablation studies with smaller sample sizes. Note that, in contrast to the standard literature on transfer learning from VLMs, which samples label-balanced adaptation sets, we follow the practice in conformal prediction [47, 46] that maintains the dataset's label-marginal distribution in the calibration data to avoid breaking exchangeability assumptions. All experiments are repeated 50 times using different random seeds for sampling the calibration data.

Zero-shot models. CLIP [39], specifically ViT-B/16 backbone, is used as a zero-shot model to evaluate the proposed conformal procedure. We also present generalization studies to other CLIP and MetaCLIP [58] backbones. The text encoder from each model is used to produce zero-shot class-wise prototypes for each downstream category by using standard templates and category names [60, 16, 45], specifically the same prompts as in [47].

Implementation details. Conformal sets are created at error rates $\alpha \in \{0.10, 0.05\}$. For T-FCP, we set $\alpha_{\text{ICP}} = 0.5\%$. For SO-LDA, the hyperparameters are fixed to $\lambda_{\text{TEXT}} = 1$ and $\lambda_{\text{REG}} = 10$.

Baselines. First, we compare our conformal procedure with standard ones in the literature:

- Inductive conformal prediction (ICP): uses the whole calibration sample for quantile search, and computes the nonconformity score using the zero-shot scores (ZS) without adaptation.
- Split conformal prediction (SCP): splits the calibration data into two equally-sized subsets, one for adaptation and the other for conformal quantile search. Since SCP is inductive, there is no computational restriction during the adaptation stage; therefore, we used standard softmax cross-entropy minimization via iterative gradient descent (GD). Training details are in Appendix E.
- Full conformal prediction (FCP): performs the full conformal procedure over the whole label space. Adaptation is performed using our proposed SO-LDA solver.

Second, we include ICP procedures that perform unsupervised transductive adaptation (ICP-T), recently proposed for VLMs [47, 46, 4]. Training details are in Appendix E.

Nonconformity scores. We use LAC [41] as the default score, as it is low-latency and commonly used in FCP [6, 35]. APS [40] is explored in Appendix F.3. More adaptive scores, such as RAPS [2] or ClusterCP [9], require sorting or additional hyperparameters, which introduces extra computation and, in some cases, additional data splits. Since our focus is on data-efficient and scalable FCP, we leave the integration of adaptive nonconformity scores into large-scale FCP to future work.

Metrics. We report accuracy and standard conformal metrics. Empirical coverage is characterized by its average and deviation (2σ) across seeds, together with the percentage of runs achieving coverage at least $1 - \alpha - 0.005$ (%Valid), being 0.005 a tolerance. Set efficiency is measured by the mean and median set size, and by the percentage of singleton predictions (%Sing.), as in [5].

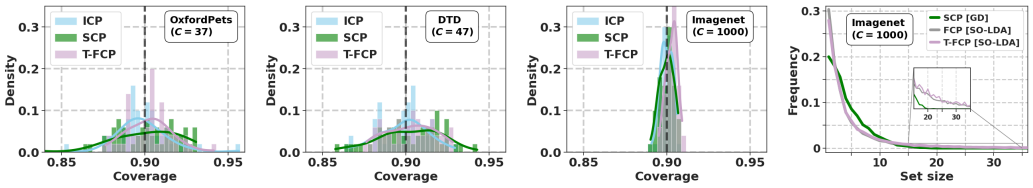
4.2 Main results

Standard conformal procedures. Table 1 provides the results obtained with the different conformal procedures, and Figure 1 (a) illustrates the empirical coverage distribution for the main conformal alternatives. (1) In terms of **coverage**, it is worth noting that, despite all procedures achieving the target average coverage, SCP's coverage deviation across samples is 20% relatively larger than ICP's, due to the data split. This is not the case for FCP procedures, which provide narrower coverage

distributions, as in ICP. Notably, T-FCP achieves the largest proportion of experiments with valid coverage among these procedures: 81.8% and 88.7% for $\alpha = 0.10$ and $\alpha = 0.05$, respectively. This is explained by the theoretically expected slightly over-coverage produced by the union-bound error rate design, i.e., α_{ICP}, for early label pruning. (2) Regarding **set efficiency**, ICP produces the largest set sizes, since it does not benefit from transfer learning with labeled data. SCP and FCP provide similar mean set sizes. However, the median indicates that FCP produces set sizes that are $\sim 10\%$ smaller (more efficient) than SCP, while its number of singletons (better on the discriminative aspect) is larger by a considerable margin. We explain these mixed observations in Figure 1 (b), showing that FCP using SO-LDA presents longer tails in the set size distribution, i.e., larger sets on uncertain inputs. This is explained by the design of the FCP, which inflates the conformity of each image-label candidate by training in the extended dataset, and the differences in optimization objectives between solvers. Finally, T-FCP provides slightly more conservative (larger) sets than FCP, with the advantage of enabling its real-time application on large-scale datasets, as discussed later. Still, T-FCP improves the median set size and singleton rate relative to SCP.

Table 1: Comparison of conformal procedures atop CLIP ViT-B/16 using $N = C \times 16$. Results are averaged across 11 datasets, and per-dataset results are in Appendix F.2.

			Cov. ($\alpha = 0.10$)			Set size			Cov. ($\alpha = 0.05$)			Set size		
		Acc.	Avg.	2σ	%Valid	Mean	Med.	%Sing.	Avg.	2σ	%Valid	Mean	Med.	%Sing.
ICP	ZS	65.7	90.1	2.0	77.8	4.8	4.5	37.3	95.2	1.5	84.0	7.9	7.3	28.2
ICP-T	OT [47]	68.6	90.6	2.0	82.3	3.9	3.7	42.4	95.4	1.4	86.8	6.1	5.6	32.6
ICP-T	TIM [46]	71.5	90.6	2.0	80.5	3.6	3.5	45.2	95.4	1.4	87.2	5.6	5.3	34.3
SCP	GD	79.9	90.7	2.7	77.1	2.2	2.1	64.0	95.5	1.9	79.5	3.3	3.0	51.9
FCP	**SO-LDA**	80.3	90.1	2.2	77.6	2.2	1.8	69.3	95.2	1.6	80.5	3.3	2.6	57.4
T-FCP	**SO-LDA**	80.3	90.7	2.1	81.8	2.3	1.9	68.1	95.6	1.5	88.7	3.5	2.7	56.0



(a) Coverage empirical distribution. (b) Set size distribution.

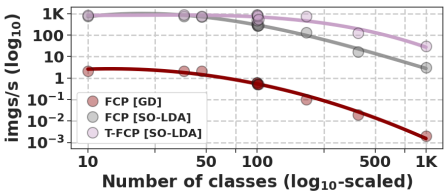
Figure 1: Analysis on empirical coverage and set size distribution on representative datasets. Results using CLIP ViT-B/16 as backbone, $N = C \times 16$ samples for calibration.

Comparison against transductive adaptation. Recently explored unsupervised transductive solvers (ICP-T in Table 1) improve over vanilla ICP in accuracy and set efficiency while maintaining the same calibration sample and therefore similar dispersion in coverage distribution. However, their unsupervised adaptation underperforms SCP in accuracy and therefore yields larger conformal sets.

Computational efficiency. Figure 2(a) showcases the latency scaling, in images per second, for full conformal procedures with the number of categories. Our experiments estimate that running the full conformal procedure using the iterative GD solver on ImageNet would require ~ 500 s per image, which is infeasible, as generally assumed in the literature. By using our SO-LDA solver, such latency is dramatically reduced to ~ 0.33 s, and is further decreased to ~ 30 ms when pruning unlikely labels via our proposed T-FCP procedure. Regarding GPU usage, we parallelized the number of tested labels in the FCP procedure to 100, which maintains, as depicted in Figure 2(b), a top-up peak-GPU memory consumption of ~ 16 Gb on the largest dataset, i.e., ImageNet. However, as shown in Figure 2(b), such a limit can be decreased without sacrificing much latency.

4.3 In-depth studies

Label pruning stage. Figure 3 (a) provides a visualization of the pruning capabilities of the inductive stage of our T-FCP procedure across datasets. Figure 3 (a, *left*) shows that most categories



(a) Latency scaling with number of classes.

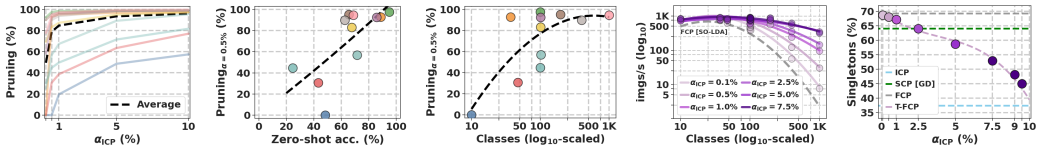
Dataset	C / Iter	P-GPU	imgs / s
DTD ($C = 37$)	25	106.3 MB	393
	50	181.0 MB	745
	100	181.1 MB	746
ImageNet ($C = 1$K)	25	4.1 GB	27.7
	50	8.2 GB	29.4
	100	16.2 GB	30.3

(b) Label parallelization (**C/Iter**) in T-FCP.

Figure 2: Computational efficiency analysis. Results using ViT-B/16, $N = C \times 16$ and $\alpha = 0.05$. All our experiments were conducted on a single NVIDIA A100-PCIE-40GB.

are discarded with small α_{ICP}, specifically the elbow is found near 1% of the allowed error rate. Therefore, most computational efficiency gains can be achieved without sacrificing much of the error-rate budget for the full conformal stage. Figure 3 (a, *center*) shows a correlation between the zero-shot performance and the pruning capabilities; thus, T-FCP will be more computationally efficient the stronger the zero-shot model initially is. However, upon a closer look to Figure 3 (a, *center* and *right*), large-scale datasets, such as SUN397 (●) or ImageNet (●), present larger pruning ratios than their zero-shot accuracy (nearly $60 - 70\%$) would imply. These results suggest that our T-FCP procedure is especially effective on the largest-scale datasets, where inter-class heterogeneities are greater and therefore categories with unrelated visual content are more easily pruned.

Conservativeness and computational feasibility trade-off. In Figure 3 (b) we explore the effect of pruning at different α_{ICP} on latency gains, as well as the set efficiency loss. As theoretically expected, larger α_{ICP} provide faster runtimes in the FCP stage according to Figure 3 (b, *left*). In terms of set efficiency, i.e, conservativeness, the produced set distribution is bounded between the ICP and FCP, as shown in Figure 3 (b, *right*), providing a user with control over the computational feasibility and set efficiency sacrifice via α_{ICP}. If the latter is a priority, then T-FCP provides gains over SCP only when using $\alpha_{\text{ICP}} < 2.5\%$. On the other hand, if coverage consistency is the main desideratum, all T-FCP configurations would yield a smaller variance in the finite-sample coverage distribution, similar to ICP, and therefore the decision on how to set α_{ICP} is constrained only by the available latency margin.



(a) Pruning control at ICP stage (results per dataset).

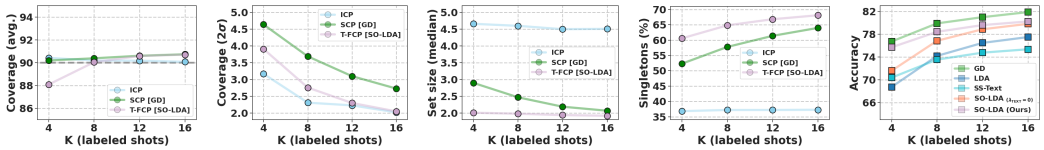
(b) Latency vs. convervativeness.

Figure 3: Ablation studies on T-FCP. Results using CLIP ViT-B/16 as backbone, $N = C \times 16$ samples for calibration, and conformal sets created at $\alpha = 0.1$ error rate.

Robustness to smaller calibration samples. Figure 4 (a-d) tests the robustness of the different conformal procedures under smaller calibration sets. Focusing on the empirical coverage dispersion Figure 4 (b), as theoretically expected, all procedures reduce it as more calibration data become available. In this context, T-FCP enables more consistent coverage than SCP across all data regimes, while providing consistently more efficient sets (Figure 4 (c-d)). In the extreme low-data regime, i.e., $N = C \times 4$, we observe a slight instability of the full conformal procedure, Figure 4 (a), which provides slightly lower marginal coverage than desired, an empirical instability already observed in prior work [35]. However, we would like to point out that the large variance of the SCP procedure ($\sim \pm 4.5$) makes its realized coverage also less stable in this regime.

SO-LDA configuration and performance. Figure 4 (e) compares the discriminative performance of the different relevant linear solvers across increasing data regimes. First, it is worth noting that methods that rely solely on class centers, e.g., Simple-shot (SS) [55, 49], despite their computational efficiency, fall short in performance compared to the best performing method, i.e., using iterative gradient-descent updates (GD), which makes them uncompetitive for full conformal procedures compared to SCP. Our proposed stabilized LDA solver with zero-shot guided feature centering in Eq. 16 consistently improves the performance of the learned prototypes compared to vanilla LDA [56].

We attribute this behavior to the fact that visual examples within a category may be highly correlated in a specific visual domain, whereas textual information can improve the estimation of the main directions of dispersion in class distributions. Notably, SO-LDA reached performances closer to the expensive iterative GD solver, just $\sim 1.5\%$ below. The constraint that learned prototypes remain close to their initial configurations in Eq. 17 helped stabilize performance gains in lower data regimes, as already observed in prior work on transfer learning from VLMs [22, 31, 45, 48].



(a) (b) (c) (d) (e)

Figure 4: Studies at lower data regimes: (a-d) conformal procedures, and (e) few-shot solvers. Results using CLIP ViT-B/16 as backbone, and conformal sets at $\alpha = 0.1$, averaged for 11 datasets.

Performance atop other VLMs. We evaluate the conformal procedures on additional backbones, including CLIP [39] and MetaCLIP [58], in Figure 5. The merits of T-FCP hold across all backbones, providing consistently more concentrated empirical coverage distributions than SCP, with variances $\sim 30\%$ smaller, while maintaining valid coverage. Moreover, predictive accuracy and set efficiency are consistently improved, especially compared to ICP. This is even the case for larger-scale backbones, e.g., MetaCLIP ViT-L/14, for which the median set size is reduced by 31%, without sacrificing finite-sample coverage stability. Despite the strong reliance of T-FCP and our SO-LDA solver on zero-shot prototypes, performance remains robust also in smaller backbones, e.g., ViT-B/32.

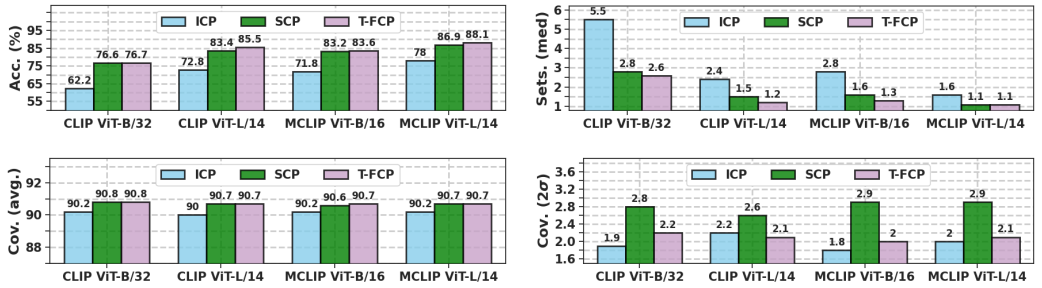


Figure 5: Generalization atop other CLIP [39] and MetaCLIP [58] backbones of conformal procedures. Results using $N = C \times 16$ samples and $\alpha = 0.1$, averaged across 11 datasets.

5 Discussion

We presented T-FCP and SO-LDA, enabling full conformal prediction on large-scale image datasets by leveraging zero-shot VLMs. Our experiments highlight that data-efficient procedures yield more stable empirical coverage than split-based alternatives and underscore the importance of data efficiency in conformal image classification.

Limitations. The proposed linear solver is still less performant than iterative gradient-based methods. Although this gap is partially compensated for in FCP by fitting on each candidate-augmented dataset, we cannot guarantee that the resulting prediction sets will always be more efficient than those from SCP, which can use more expensive solvers during its adaptation stage. A further limitation is that our main SO-LDA implementation uses diagonal loading to stabilize the precision matrix, thereby approximating recomputing a fully regularized LDA estimator for each augmented dataset. Therefore, although T-FCP offers theoretically valid coverage, the same is not true when it is combined with our SO-LDA solver. In practice, ablations in Appendix C show negligible differences against exact implementations. Finally, the coverage guarantee of T-FCP is based on an additive union-bound allocation across the pruning and full-conformal stages. Consequently, the resulting sets may be slightly more conservative than FCP's. Hence, FCP may be preferred for smaller-scale tasks, e.g., those involving tens or a few hundred classes, especially when coupled with our online LDA solver.

Disclosure of Interests. The authors have no conflicts of interest or funding to disclose that are relevant to the content of this article.

References

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

[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*, pages 1–17, 2020.

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

[4] Behzad Bozorgtabar, Dwarikanath Mahapatra, Sudipta Roy, Imran Razzak Muzammal Naseer, and Zongyuan Ge. LATA: Laplacian-Assisted Transductive Adaptation for Conformal Uncertainty in Medical VLMs . In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2026.

[5] Margarida M Campos, João Cálem, Sophia Sklaviadis, Mario A. T. Figueiredo, and Andre Martins. Sparse activations as conformal predictors. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan, editors, *Proceedings of International Conference on Artificial Intelligence and Statistics*, volume 258, pages 2674–2682, 2025.

[6] Giovanni Cherubin, Konstantinos Chatzikokolakis, and Martin Jaggi. Exact optimization of conformal predictors via incremental and decremental learning. In *International Conference on Machine Learning*, pages 1836–1845. PMLR, 2021.

[7] 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*, pages 3606–3613, 2014.

[8] 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*, pages 248–255, 2009.

[9] 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*, volume 36, pages 64555–64576, 2023.

[10] Tiffany Ding, Jean-Baptiste Fermanian, and Joseph Salmon. Conformal prediction for long-tailed classification. In *The Fourteenth International Conference on Learning Representations*, 2026.

[11] 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.

[12] 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*, pages 178–178, 2004.

[13] Leo Fillioux, Julio Silva-Rodriguez, Ismail Ben Ayed, Paul-Henry Cournede, Maria Vakalopoulou, Stergios Christodoulidis, and Jose Dolz. Are foundation models for computer vision good conformal predictors? *Transactions on Machine Learning Research*, 2026.

[14] Adam Fisch, Tal Schuster, Tommi Jaakkola, and Regina Barzilay. Efficient conformal prediction via cascaded inference with expanded admission. In *Proceedings of The Tenth International Conference on Learning Representations*, 2021.

[15] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

[16] 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*, 2023.

[17] Etash Kumar Guha, Eugene Ndiaye, and Xiaoming Huo. Conformalization of sparse generalized linear models. In *International Conference on Machine Learning*, pages 11871–11887. PMLR, 2023.

[18] William W Hager. Updating the inverse of a matrix. *SIAM Review*, 31(2):221–239, 1989.

[19] 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*, pages 3606–3613, 2018.

[20] 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*, volume 235, pages 20331–20347, 2024.

[21] Roel Hulsman. Distribution-free finite-sample guarantees and split conformal prediction. *arXiv preprint arXiv:2210.14735*, 2022.

[22] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023.

[23] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013.

[24] Tatsuya Kubokawa and Muni S Srivastava. Estimation of the precision matrix of a singular wishart distribution and its application in high-dimensional data. *Journal of multivariate Analysis*, 99(9):1906–1928, 2008.

[25] Ludmila I. Kuncheva and Catrin O. Plumpton. Adaptive learning rate for online linear discriminant classifiers. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 510–519, 2008.

[26] Jing Lei. Fast exact conformalization of the lasso using piecewise linear homotopy. *Biometrika*, 106(4):749–764, 2019.

[27] 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.

[28] Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

[29] Diyang Li. Generalized fast exact conformalization. In *Neural Information Processing Systems*, 2024.

[30] 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.

[31] 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*, pages 19325–19337, 2023.

[32] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen AWM van der Laak, Bram van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

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

[34] F Marques and C Paulo. Universal distribution of the empirical coverage in split conformal prediction. *Statistics & Probability Letters*, 219(C), 2025.

[35] Javier Abad Martinez, Umang Bhatt, Adrian Weller, and Giovanni Cherubin. Approximating full conformal prediction at scale via influence functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6631–6639, 2023.

[36] 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.

[37] 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.

[38] 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*, page 3498–3505, 2012.

[39] 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*, pages 8748–8763, 2021.

[40] Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, pages 3581–3591, 2020.

[41] Mauricio Sadinle, Jing Lei, and Larry Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

[42] Sara Sangalli, Gary Sarwin, Ertunc Erdil, Carlo Serra, Alessandro Carretta, Victor Staartjes, and Ender Konukoglu. Conformal forecasting for surgical instrument trajectory . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15968. Springer Nature Switzerland, September 2025.

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

[44] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

[45] 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*, pages 23681–23690, 2024.

[46] Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz. Trustworthy Few-Shot Transfer of Medical VLMs through Split Conformal Prediction . In *Proceedings of Medical Image Computing and Computer Assisted Intervention*, volume LNCS 15966, pages 658–668, 2025.

[47] Julio Silva-Rodríguez, Ismail Ben Ayed, and Jose Dolz. Conformal prediction for zero-shot models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19931–19941, 2025.

[48] Julio Silva-Rodríguez et al. Few-shot, now for real: Medical vlms adaptation without balanced sets or validation. In *Proceedings of Medical Image Computing and Computer Assisted Intervention*, volume LNCS 15966, 2025.

[49] Julio Silva-Rodríguez et al. Full conformal adaptation of medical vision-language models. In *International Conference on Information Processing in Medical Imaging*, pages 278–293, 2025.

[50] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. In *arXiv preprint arXiv:1212.0402*, 2012.

[51] David Stutz, Krishnamurthy Dj Dvijotham, Ali Taylan Cemgil, and Arnaud Doucet. Learning optimal conformal classifiers. In *International Conference on Learning Representations (ICLR)*, 2022.

[52] Alexander Timans, Christoph-Nikolas Straehle, Kaspar Sakmann, and Eric Nalisnick. Adaptive bounding box uncertainties via two-step conformal prediction. In *European Conference on Computer Vision*, pages 363–398. Springer, 2024.

[53] Vladimir Vovk. Conditional validity of inductive conformal predictors. In *Proceedings of the Asian Conference on Machine Learning*, volume 25, pages 475–490, 2012.

[54] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 01 2005.

[55] Yan Wang et al. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. In *arXiv preprint arXiv:1911.04623*, 2019.

[56] Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A hard-to-beat baseline for training-free CLIP-based adaptation. In *International Conference on Learning Representations*, 2024.

[57] 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*, pages 3485–3492, 2010.

[58] 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*, 2024.

[59] 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*, pages 1–19, 2022.

[60] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2022.

A Broader Impacts

Broader impacts. Conformal prediction is often motivated by providing theoretical guarantees in high-stakes deployment. However, it should be interpreted with care. Marginal coverage does not imply conditional or subgroup coverage, and guarantees need not hold under distribution shift. Moreover, finite-sample empirical coverage can fall below the nominal level in a real-world deployment, especially with data-inefficient pipelines. In this work, we directly address the latter use case, enabling more stable conformal image classifiers via full conformal prediction. We do not anticipate negative societal impacts from our work beyond those of standard conformal prediction, and we identify no domain-specific risks requiring special attention.

B Proofs

Proof of Proposition 1. According to Eq 9:

$$\mathbb{P}\big(Y \notin \mathcal{C}_{1\cap 2}(\mathbf{x})\big) = \mathbb{P}\big(Y \notin \mathcal{C}_1(\mathbf{x}) \ \cup \ Y \notin \mathcal{C}_2(\mathbf{x})\big). \tag{18}$$

By the union bound,

$$\mathbb{P}\big(Y \notin \mathcal{C}_{1\cap 2}(\mathbf{x})\big) \leq \mathbb{P}\big(Y \notin \mathcal{C}_1(\mathbf{x})\big) + \mathbb{P}\big(Y \notin \mathcal{C}_2(\mathbf{x})\big). \tag{19}$$

Using the marginal coverage guarantees for each individual procedure in Eq. 8,

$$\mathbb{P}\big(Y \notin \mathcal{C}_{1\cap 2}(\mathbf{x})\big) \leq \alpha_1 + \alpha_2, \tag{20}$$

which implies the result. □

C Stability of SO-LDA

In this section, we further analyze SO-LDA, particularly the implications of the mild asymmetry introduced by the diagonal covariance matrix regularization used in the main implementation, whose diagonal is estimated only from calibration data to enable efficient rank-one updates. We compare such a solution with different configurations:

- **Ridge**: the regularization is imposed via a fixed identity loading, $\mathbf{S}_{(N)}^{\text{reg}} = \mathbf{S}_{(N)} + \lambda_{\text{RIDGE}}\mathbf{I}$, which is compatible with permutation-invariant rank-one updates under fixed regularization.
- **Non-online**: for each candidate sample-label pair, it recomputes the class means and the covariance matrix from the full augmented dataset, then directly inverts the regularized covariance matrix. This solver increases the per-label computational cost to $\mathcal{O}(NF^2 + F^3)$.
- **Unstable**: does not use zero-shot prototypes for the residuals estimation as in Eq. 16. Hence, the centered features are calculated asymmetrically between calibration and test data.

Comparative performance on different data regimes is presented in Table 2. The results suggest that the dominant source of instability is the asymmetric centering of the residuals used for covariance estimation, with the **Unstable** configuration yielding the largest marginal coverage gap and lower ratio of valid occurrences. The fixed-**Ridge** configuration is the cleanest variant with respect to symmetric online updates. However, its discriminative performance is lower than that of our diagonal regularization in SO-LDA. Moreover, both exhibit a similar coverage distribution, suggesting that the asymmetry introduced by our implementation choice is negligible. Also, the performance levels achieved by SO-LDA are similar to those obtained with more expensive, exact **Non-online** updates.

The under-coverage for $K = 4$ is observed in both online and exact recomputation variants, suggesting it is not primarily caused by the diagonal-loading approximation. As this instability in exact FCP also appears in prior work [35] (e.g., see their results at their Figure 4(d)), we leave a complete characterization of this regime to future work.

D Datasets

A summary of the datasets, number of classes, tasks, and number of samples in the used test partition is depicted in Table 3.

Table 2: Study of the SO-LDA solver configurations under T-FCP. Average results for 10 datasets. In contrast to the main manuscript, these studies use 20 random seeds and exclude ImageNet.

	Cov. ($\alpha = 0.10$)				Set size	
	Acc.	Avg.	2σ	%Valid	Mean	%Sing.
SO-LDA	81.0	90.8	2.1	80.0	2.3	70.3
Non-online	81.0	90.8	2.1	80.0	2.3	70.2
Ridge	78.8	90.7	2.3	76.0	2.4	64.6
Unstable	78.8	90.4	2.3	67.5	2.6	62.3

(a) $N = C \times 16$

	Cov. ($\alpha = 0.10$)				Set size	
	Acc.	Avg.	2σ	%Valid	Mean	%Sing.
SO-LDA	76.4	88.2	4.2	27.0	3.1	61.6
Non-online	76.4	88.3	4.2	27.5	3.2	61.3
Ridge	74.1	88.0	4.3	26.0	2.9	58.4
Unstable	70.1	86.0	5.3	13.0	4.6	45.9

(b) $N = C \times 4$

Table 3: Datasets overview. "b"/"i" denotes balanced/imbalanced.

Dataset	Classes	Splits			b/i	Task description
		Train	Val	Test		
EuroSAT [19]	10	13,500	5,400	8,100	i	Satellite image classification.
OxfordPets [38]	37	2,944	736	3,669	i	Pets classification.
DTD [7]	47	2,820	1,128	1,692	b	Textures classification.
FGVCAircraft [33]	100	3,334	3,333	3,333	i	Aircraft classification.
Caltech101 [12]	100	4,128	1,649	2,465	i	Natural objects classification.
UCF101 [50]	101	7,639	1,898	3,783	i	Action recognition.
Food101 [3]	101	50,500	20,200	30,300	b	Foods classification.
Flowers102 [36]	102	4,093	1,633	2,463	u	Flowers classification.
StanfordCars [23]	196	6,509	1,635	8,041	i	Cars classification.
SUN397 [57]	397	15,880	3,970	19,850	b	Scenes classification.
ImageNet [8]	1,000	1.28M	-	50,000	b	Natural objects recognition.

E Baselines

Gradient-descent linear probe (GD). We followed training details similar to [45], which remain fixed for all datasets. The learned class weights are initialized with zero-shot prototypes. Full-batch gradient descent is performed for 300 iterations, minimizing cross-entropy. SGD is used as the optimizer with momentum 0.9. A cosine-annealing scheduler with a base learning rate of 0.1 is applied to ensure a small step size and proper convergence.

ICP-T via Conf-OT [47]. The logits are extracted using the zero-shot prototypes, and label-marginal distributions are estimated from calibration data. The Optimal Transport solver is applied to the joint calibration and test sets, with the entropic weight $\tau = 1$ and $T_{\text{OT}} = 3$, as default values in [47].

ICP-T via TIM [46]. We use the Transductive Information Maximization solver with regularization on the expected label-marginal distribution, estimated from calibration labels. Training is carried out on the joint calibration and test samples. The relative weight λ is set to 1.0, as in [46]. Similarly to GD, full-batch gradient descent is performed for 300 iterations, in this case optimizing the mutual information criteria. Other details are as in GD, since they provided satisfactory convergence.

F Detailed Results

F.1 Additional configurations

We report the performance of the proposed solver SO-LDA for split conformal prediction in Table 4.

Table 4: Comparison of conformal procedures atop CLIP ViT-B/16 using $N = C \times 16$. This table reports the results using configurations not included in Table 1.

		Cov. ($\alpha = 0.10$)			Set size			Cov. ($\alpha = 0.05$)			Set size		
	Acc.	Avg.	2σ	%Valid	Mean	Med.	%Sing.	Avg.	2σ	%Valid	Mean	Med.	%Sing.
SCP SO-LDA	79.5	90.8	3.5	71.3	2.7	2.3	69.4	95.6	2.4	77.0	4.1	3.3	58.4

F.2 Table 1: detailed results per dataset

The detailed results of the conformal procedures' performance across datasets are reported in Table 5.

Table 5: Per-dataset conformal prediction results of average performances of the main conformal procedures reported in Table 1, using CLIP ViT-B/16.

				Cov. ($\alpha = 0.10$)			Set size			Cov. ($\alpha = 0.05$)			Set size		
			Acc.	Avg.	2σ	%Valid	Mean	Med.	%Sing.	Avg.	2σ	%Valid	Mean	Med.	%Sing.
EuroSAT	ICP	ZS	48.4	90.4	6.1	62.0	4.2	4.4	9.4	95.6	4.8	66.0	5.2	5.3	6.6
	SCP	GD	83.4	89.8	6.5	50.0	1.4	1.1	73.4	96.1	4.7	76.0	2.2	1.8	43.8
	T-FCP	**SO-LDA**	83.0	90.2	5.3	58.0	1.4	1.0	71.4	95.9	3.9	90.0	2.0	1.9	42.0
Pets	ICP	ZS	88.8	89.9	3.2	54.0	1.0	1.0	92.9	95.1	2.0	68.0	1.2	1.0	80.4
	SCP	GD	92.3	90.0	4.9	65.0	1.0	1.0	94.8	95.2	3.3	66.0	1.1	1.0	91.0
	T-FCP	**SO-LDA**	92.8	90.3	2.7	70.0	1.0	1.0	93.6	95.6	2.2	82.0	1.1	1.0	91.0
DTD	ICP	ZS	43.4	89.9	2.9	66.0	10.8	11.1	6.75	95.0	1.9	72.0	17.2	18.6	4.3
	SCP	GD	66.6	90.2	4.4	55.0	3.4	2.8	28.3	95.1	2.6	72.0	6.2	5.3	16.7
	T-FCP	**SO-LDA**	69.7	90.3	2.9	66.0	3.2	2.0	43.2	95.0	2.0	70.0	6.1	3.4	29.7
Aircraft	ICP	ZS	24.8	90.0	1.3	76.0	18.4	18.1	0.8	95.2	1.2	90.0	28.6	28.0	0.1
	SCP	GD	41.4	90.1	2.3	65.0	8.8	8.8	3.7	95.0	1.8	70.0	12.7	12.7	2.1
	T-FCP	**SO-LDA**	39.6	90.2	2.1	70.0	9.4	8.9	7.2	95.1	1.5	78.0	13.1	12.5	5.9
Caltech	ICP	ZS	95.1	90.4	1.6	88.0	0.9	1.0	93.4	96.2	1.0	100.0	1.0	1.0	94.0
	SCP	GD	97.8	94.9	1.4	100.0	1.0	1.0	95.5	97.8	0.9	100.0	1.0	1.0	99.0
	T-FCP	**SO-LDA**	97.9	94.1	0.9	100.0	1.0	1.0	94.7	97.3	0.9	100.0	1.0	1.0	98.4
UCF101	ICP	ZS	67.6	90.1	1.7	74.0	2.9	2.0	29.3	95.1	1.4	80.0	5.1	3.9	16.4
	SCP	GD	89.3	90.3	3.0	70.0	1.0	1.0	90.5	95.1	1.8	68.0	1.3	1.0	75.1
	T-FCP	**SO-LDA**	89.6	90.5	1.9	86.0	1.0	1.0	94.8	95.4	1.4	86.0	1.3	1.0	79.2
Food101	ICP	ZS	86.0	90.1	2.0	72.0	1.1	1.0	85.1	95.1	1.4	72.0	1.6	1.0	63.4
	SCP	GD	86.3	90.0	2.9	60.0	1.1	1.0	86.8	94.8	2.0	60.0	1.5	1.0	66.5
	T-FCP	**SO-LDA**	86.4	90.4	1.9	80.0	1.2	1.0	87.8	95.4	1.6	90.0	1.7	1.0	68.3
Flowers	ICP	ZS	71.5	90.3	0.8	98.0	4.9	3.9	19.7	95.2	0.8	96.0	11.1	9.4	5.67
	SCP	GD	96.5	92.2	2.1	100.0	0.9	1.0	93.0	96.3	1.4	98.0	1.0	1.0	96.9
	T-FCP	**SO-LDA**	95.7	90.9	1.6	96.0	0.9	1.0	92.9	95.8	1.1	100.0	1.0	1.0	96.1
Cars	ICP	ZS	65.5	90.1	1.0	86.0	2.4	2.0	27.6	95.2	0.9	88.0	3.4	3.0	17.3
	SCP	GD	78.6	90.3	1.8	90.0	1.5	1.0	59.3	95.2	1.1	86.0	2.0	2.0	38.3
	T-FCP	**SO-LDA**	80.8	90.4	1.2	90.0	1.4	1.0	65.4	95.3	0.9	96.0	1.9	2.0	45.9
SUN397	ICP	ZS	62.6	90.2	1.0	90.0	3.5	3.0	16.2	95.1	0.7	98.0	6.6	5.3	7.1
	SCP	GD	73.7	90.0	1.5	80.0	1.9	2.0	40.8	95.1	1.1	84.0	3.2	2.9	22.0
	T-FCP	**SO-LDA**	74.3	90.3	1.4	84.0	2.0	1.2	51.3	95.3	1.0	94.0	3.4	2.0	31.2
ImageNet	ICP	ZS	68.7	90.0	0.7	90.0	2.8	2.0	28.9	95.0	0.6	94.0	5.5	4.0	15.0
	SCP	GD	72.5	90.0	0.8	85.0	2.3	2.0	36.9	95.0	0.5	94.0	4.2	3.0	20.1
	T-FCP	**SO-LDA**	72.9	90.3	0.7	100.0	2.5	2.0	46.7	95.4	0.4	100.0	5.7	3.0	27.9

F.3 Results using APS

We report results using APS [40] as the non-conformity score in Table 6. They demonstrate that the gains in empirical coverage stability are not specific to LAC, but also generalize to APS: T-FCP reports smaller coverage dispersion and a larger ratio of realizations with satisfied coverage than ICP and SCP. We also incorporate class-conditional average coverage gap (CCV), the reference metric used for measuring adaptiveness [9], where T-FCP remains consistent.

It is worth noting that adaptive scores such as APS/RAPS require ranking or sorting over the label space. This adds non-negligible cost inside the candidate-wise FCP loop, which conflicts with our scalability focus. For example, for ImageNet, T-FCP [$\alpha_{\text{ICP}} = 0.5\%,\ \alpha_{\text{FCP}} = 9.5\%$] using LAC has a latency of nearly 30 ms/image, and it increases to nearly 220 ms/image using APS, which makes it currently less appealing to use for the largest-scale datasets for T-FCP and unfeasible for the more expensive vanilla FCP.

Table 6: Results using APS. SO-LDA is used as solver, with $N = C \times 16$. Average results for 10 datasets. These studies use 20 random seeds and exclude ImageNet.

	Cov. ($\alpha = 0.10$)				Set size			Cov. ($\alpha = 0.05$)				Set size		
	Avg.	2σ	%Valid	CCV	Mean	Med.	%Sing.	Avg.	2σ	%Valid	CCV	Mean	Med.	%Sing.
ICP	90.4	2.2	81.5	8.2	6.3	4.9	34.4	95.3	1.5	85.0	5.0	10.1	8.0	27.1
SCP	90.1	3.1	67.5	6.3	3.0	2.2	54.0	95.1	2.4	70.0	5.0	4.3	3.1	46.9
T-FCP	90.4	1.7	84.0	6.4	3.3	2.3	52.6	95.3	1.4	89.0	4.9	4.5	2.9	46.9

F.4 Extension to unimodal models

The T-FCP construction is model-agnostic in principle: any valid low-cost conformal predictor could be used as the pruning stage, and any valid or efficient FCP scoring rule could be used as the second stage. Beyond VLMs, T-FCP can therefore be explored for unimodal models as long as an initial pruning proxy exists, e.g., a pretrained classifier over the target label space.

To test this, we next transfer a standard ResNet-50 model pretrained on ImageNet to ImageNet-V2 (accuracy 69.1%), since the two datasets share the same label space. T-FCP [$\alpha_{\text{ICP}} = 0.5\%,\ \alpha_{\text{FCP}} = 4.5\%$] early prunes 65% of the labels, and provides a more stable empirical coverage distribution than split-conformal alternatives. Specific results are presented below, in Table 7.

Table 7: Conformal prediction results on ImageNet-V2 using ResNet-50 pre-trained on ImageNet, LAC, and 8 labeled samples per class as calibration data. Results across 20 seeds.

	Cov. ($\alpha = 0.05$)			Set size
	Avg.	2σ	%Valid	Mean
ICP	95.2	1.3	80.0	12.3
SCP	95.0	1.2	70.0	11.7
T-FCP	95.1	1.2	80.0	10.6

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims are appropriately justified in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: All assumptions and limitations were stated where appropriate.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Proofs can be found in the supplementary material, and all assumptions were appropriately stated in the main manuscript.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Implementation details, including hyperparameters, are detailed. The experiments are carried out using public datasets and models.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

Justification: The code will be publicized upon acceptance of the manuscript, and is shared for review.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Data splits and hyperparameters are indicated in the main manuscript.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not provide confidence intervals, but results are averaged across randomized samples, and variability, via standard deviation, is reported for the relevant metrics.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: GPU memory usage, latency, and resources are indicated in the main manuscript and supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The topic is not particularly sensitive, and all data comes from open datasets. There are no ethics violations.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed the Broader Impacts in the supplementary material.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point

out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Given the lack of immediate negative impacts, we do not see a need for safeguards for our work.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets are open and were appropriately cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [No]

Justification: No new assets were released.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: No new assets were released.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: There was no research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: No significant use of LLMs beyond writing and editing.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.