Conformal Prediction Meets Long-tail Classification

Shuqi Liu, Jianguo Huang, Luke Ong

College of Computing and Data Science, Nanyang Technological University, Singapore
shuqi005@e.ntu.edu.sg, {jianguo.huang, luke.ong}@ntu.edu.sg

Abstract

Conformal Prediction (CP) is a popular method for uncertainty quantification that converts a pretrained model's point prediction into a prediction set, with the set size reflecting the model's confidence. Although existing CP methods are guaranteed to achieve marginal coverage, they often exhibit imbalanced coverage across classes under long-tailed label distributions, tending to over cover the head classes at the expense of under covering the remaining tail classes. This under coverage is particularly concerning, as it undermines the reliability of the prediction sets for minority classes, even with coverage ensured on average. In this paper, we propose the Tail-Aware Conformal Prediction (TACP) method to mitigate the under coverage of the tail classes by utilizing the long-tailed structure and narrowing the head-tail coverage gap. Theoretical analysis shows that it consistently achieves a smaller head-tail coverage gap than standard methods. To further improve coverage balance across all classes, we introduce an extension of TACP: soft TACP (sTACP) via a reweighting mechanism. The proposed framework can be combined with various non-conformity scores, and experiments on multiple long-tailed benchmark datasets demonstrate the effectiveness of our methods.

1 Introduction

Uncertainty quantification (UQ) is crucial in building reliable machine learning systems, especially in high-stakes applications, such as autonomous driving (Grigorescu et al. 2020), where overconfident errors can be costly. A popular approach for UQ is Conformal Prediction (CP) (Vovk, Gammerman, and Shafer 2005), which converts the point predictions of a pretrained model into prediction sets that come with rigorous statistical guarantees. Specifically, CP ensures that the prediction set contains the true label with a user-specified probability on average, and the size of the prediction set serves as a proxy for model confidence: smaller sets imply higher confidence.

Despite their strong theoretical guarantees, standard CP methods often struggle under long-tailed (LT) label distributions (Buda, Maki, and Mazurowski 2018), which are common in real-world applications, leading to imbalanced coverage across classes. In this LT setting, a small number of *head* classes dominate the data, while many other *tail* classes have few samples. CP methods typically achieve the desired marginal coverage on average, but display specific coverage imbalance: over covering the head classes at the cost of under covering the tail classes (Löfström et al. 2015; Lu et al. 2022). Conformal prediction sets often fail to include the ground-truth tail label with the target probability, resulting in unreliable uncertainty estimations. Such under coverage can mislead subsequent downstream decision-making, potentially causing costly or even harmful outcomes in fairness-sensitive or safety-critical scenarios.

To address this coverage imbalance observed in standard CP, group-conditional and class-conditional approaches have been proposed (Vovk 2013; Ding et al. 2023; Shi et al. 2024), which aim to achieve coverage guarantee for each group or individual class. However, these methods are often less practical when the calibration data is limited and exhibits a long-tailed distribution, as the per-group or per-class acceptance thresholds estimates become highly unreliable and lead to excessively large prediction sets. These limitations highlight the need for conformal prediction methods that provide valid, efficient, and balanced uncertainty estimates between groups with head and tail classes or further across classes in LT settings.

To tackle the problems above, we propose the **T**ail-**A**ware **C**onformal **P**rediction (TACP) method, which introduces a tailored regularization term that adapts to the underlying label imbalance, leading to more balanced coverage between head and tail classes. Then we provide a theoretical analysis demonstrating that TACP narrows the coverage gap between head and tail classes compared with standard CP. To further improve coverage balance across all individual classes, we introduced an extension of TACP: **s**oft **T**ail-**A**ware **C**onformal **P**rediction (sTACP), which employs a smooth reweighting strategy for adaptive penalty control. Our proposed frameworks are flexible and can be combined with a wide range of representative non-conformity scores.

Extensive experiments on multiple long-tailed benchmarks, including CIFAR100-LT and ImageNet-LT, demonstrate the effectiveness of the proposed TACP and sTACP frameworks. TACP significantly reduces the head-tail coverage gap compared to standard CP across four non-conformity scores, while maintaining informative prediction sets. For example, on ImageNet-LT ($\rho = 0.6$) using the APS

score, TACP reduces CovGap-HT from 2.18 to 1.11, accompanied by a decrease in AvgSize (36.43 $\rightarrow$ 33.98). sTACP further improves class-conditional coverage balance, outperforming other baselines on ImageNet-LT. For instance, with APS as the base score at $\alpha = 10\%$, sTACP reduces the class-conditional gap from 19.00% (STANDARD) to 15.86% while preserving a similar set size. These results highlight that TACP and sTACP provide robust, efficient, and balance uncertainty estimates under long-tailed settings.

We summarize our contributions as follows:

- We study the problem of CP under LT label distributions and propose the TACP framework to mitigate the coverage gap between head and tail classes.
- We provide a theoretical analysis to demonstrate the effectiveness of TACP and empirically compare its performance with several baselines across multiple long-tailed benchmarks using four non-conformity scores.
- To narrow class-conditional coverage gaps in LT scenarios, we propose sTACP, an enhanced variant of TACP. Extensive experiments demonstrate that sTACP consistently reduces the class-conditional coverage gap across a wide range of long-tailed classification tasks.

2 Related Work

Conformal prediction is an active area of research that provides a model-free and distribution-free framework for UQ by converting the point prediction of any model into a prediction set that contains the ground-truth label with a user-defined level. A widely used variant of CP is split CP (Papadopoulos et al. 2002; Lei et al. 2018), which employs a held-out calibration set to improve the computational efficiency of full CP (Vovk, Gammerman, and Shafer 2005). In addition, several variants of CP have been developed based on cross-validation (Vovk 2015) or leave-one-out (Barber et al. 2021; Lee and Zhang 2025). There is an increasing amount of research focused on maintaining marginal coverage by relaxing the exchangeability assumption to account for covariate shift (Tibshirani et al. 2019) and label shift (Podkopaev and Ramdas 2021; Xu et al. 2025).

CP has been applied to various tasks, including classification (Lei, Robins, and Wasserman 2013; Romano, Sesia, and Candès 2020; Angelopoulos et al. 2021; Zhou and Sesia 2024; Caprio et al. 2025; Caprio 2025), regression (Romano, Patterson, and Candes 2019; Sesia and Romano 2021; Seedat et al. 2023), and outlier detection (Guan and Tibshirani 2022; Bates et al. 2023). Its performance is typically assessed by coverage validity and length efficiency. Hence, a central research goal is to improve length efficiency (Huang et al. 2024; Xi et al. 2025; Liu et al. 2024). Another important line of research seeks to move beyond marginal coverage guarantees, including methods that aim to improve class-conditional coverage (Shi, Ong, and Leckie 2013; Sadinle, Lei, and Wasserman 2019; Plassier et al. 2025), feature-conditional coverage (Romano, Sesia, and Candès 2020). Such conditional guarantees are especially important in LT data settings, where marginal coverage can obscure significant disparities across different classes or subgroups (Löfström et al. 2015; Kasa and Taylor 2023).

Label distributions in real-world applications are often long-tailed (Buda, Maki, and Mazurowski 2018), where a few head classes dominate the data while many tail classes have a limited number of samples. Under such LT settings, previous works focused on improving models' prediction accuracy by data re-sampling (Chawla et al. 2002; Brodersen et al. 2010) and surrogate loss function modification (Cao et al. 2019a; Tan et al. 2020; Cui et al. 2019).

Previous work has observed the phenomenon of imbalanced coverage when applying conformal prediction to long-tail or imbalanced label distributions (Löfström et al. 2015; Lu et al. 2022). Kasa and Taylor (2023) empirically showed that performance degrades in LT regimes and highlighted the coverage imbalance across classes. Moreover, Shi et al. (2024) considered a partial LT setting, where the training dataset is highly imbalanced, while the calibration and test datasets used for CP remain balanced. In contrast, we consider a fully long-tailed setting, where the training, calibration, and test datasets all follow long-tailed label distributions. This setting poses challenges for achieving reliable coverage guarantees across classes, and CP methods tailored for this scenario remain relatively undeveloped.

3 Preliminaries

Conformal Prediction

In conformal prediction, let the feature space be $\mathcal{X} \subseteq \mathbb{R}^d$ and the label space be $\mathcal{Y} = \{1, 2, \ldots, K\}$ with K classes. We focus on the feature-label random variable $X \times Y \in \mathcal{X} \times \mathcal{Y}$, which has a joint distribution with density $p(\boldsymbol{x}, y)$ where $\boldsymbol{x}, y$ are their realizations. The calibration samples $\{(X_i, Y_i)\}_{i=1}^n$ and the test point (X_{n+1}, Y_{n+1}) are independently and identically drawn (i.i.d.) from the unknown distribution $p(\boldsymbol{x}, y)$ (the i.i.d. assumption can be relaxed to an exchangeability assumption (Vovk, Gammerman, and Shafer 2005)). Ordinary CP in classification tasks aims at constructing a set predictor $\mathcal{C} : \mathcal{X} \rightarrow 2^{\mathcal{Y}}$ such that for a new test point (X_{n+1}, Y_{n+1}), the following coverage guarantee holds:

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1})) \geq 1 - \alpha, \tag{1}$$

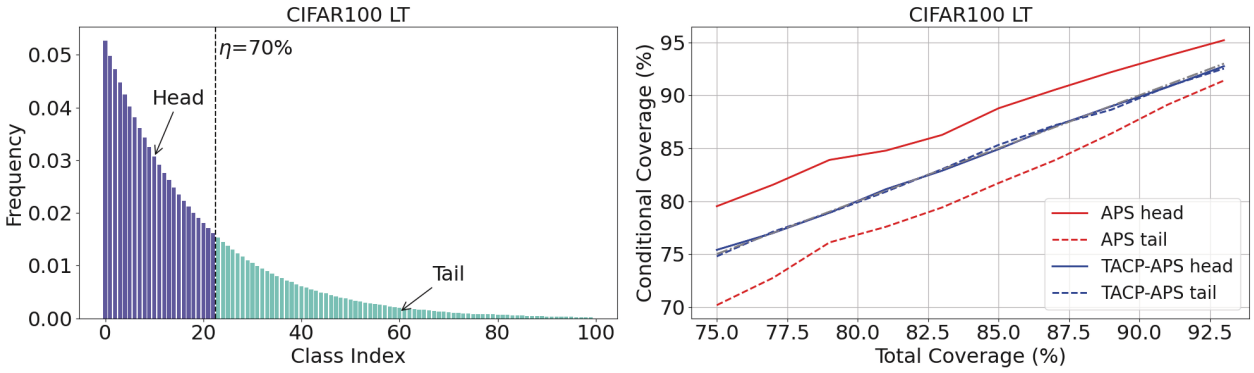
where $\alpha \in (0, 1)$ is a predefined miscoverage level. The typical approach for generating such a prediction set is to design a non-conformity score function $s : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$ that quantifies the uncertainty of the pretrained model $\hat{\pi} : \mathcal{X} \rightarrow \Delta^{K-1}$ and then return a label set whose scores fall below some acceptance threshold. For example, split conformal prediction takes the acceptance threshold as the $1 - \alpha$ quantile of the calibration non-conformity scores:

$$\mathcal{C}(X_{n+1}) = \{y : s(X_{n+1}, y) \leq \hat{\tau}_\alpha\}, \tag{2}$$

where $\hat{\tau}_\alpha = \text{Quantile}\,(1 - \alpha, \{s(X_i, Y_i)\}_{i=1}^n)$. We can make the prediction set more conservative by defining $\hat{\tau}_\alpha$ as the $\lceil (n+1)(1-\alpha) \rceil / n$ quantile and refer to this baseline as STANDARD. CP is a useful approach that provides a marginal coverage guarantee, i.e., Eq. (1), without requiring any assumptions on the model or data distribution.

Class-Conditional Coverage

However, marginal coverage only ensures that predictions are valid on average over the entire population, which often



(a) (b)

Figure 1: Left: Head-tail partitioning at $\eta = 70\%$. Right: Head- and tail-conditional coverage as a function of total coverage for the STANDARD and TACP methods using APS non-conformity score. Both figures are based on CIFAR100 LT with an imbalance factor $\mu = 100$, which quantifies the ratio between the most and least frequent classes.

masks imbalanced errors across subgroups (Löfström et al. 2015; Ding et al. 2023). To address this, a stronger notion is class-conditional coverage, which imposes the coverage guarantee for each class $y \in \mathcal{Y}$ individually:

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}(X_{n+1}) \mid Y_{n+1} = y) \geq 1 - \alpha. \quad (3)$$

A direct approach to achieve Eq. (3) is to split the calibration dataset by class and applying conformal prediction separately to each class (Vovk 2013), denoted by CLASSWISE. However, the per-class acceptance threshold estimators in CLASSWISE can be highly noisy for classes with limited samples, often causing overly large prediction sets (Shi, Ong, and Leckie 2013) (See Exp.3 in Section 7).

To mitigate the inefficiency of prediction sets, Ding et al. (2023) introduced CLUSTER, which groups "similar" classes by clustering based on quantiles of their non-conformity score distributions. Calibration samples are then assigned to these clusters, and standard CP is applied within each cluster. Moreover, Shi et al. (2024) provided the Rank Calibrated Class-conditional CP (RC3P) method, which aims to achieve class-conditional coverage while improving prediction efficiency by calibrating the acceptance thresholds for each class using a label-rank calibration procedure.

Although CLUSTER and RC3P methods significantly improve efficiency while approximating class-conditional coverage, both methods require splitting the calibration set by clusters or classes and estimating acceptance thresholds within each subset. In the LT setting, the scarcity of tail-class samples leads to highly unreliable threshold estimates, thereby limiting the practicality of such methods. This challenge is even more pronounced in the fully long-tailed setting we consider, where the training, calibration, and test sets all follow long-tailed label distributions.

4 Methods

CP under LT Distributions

Head-Tail Partition In LT settings, a common approach is to partition the label space $\mathcal{Y}$ into *head* classes $\mathcal{G}_h$ and *tail* classes $\mathcal{G}_t$. Specifically, we define the head classes $\mathcal{G}_h$ as the smallest subset of labels whose cumulative class prior probability exceeds a pre-defined threshold $\eta \in (0, 1)$:

$$\mathcal{G}_h = \underset{\mathcal{G} \subseteq \mathcal{Y}}{\operatorname{argmin}} |\mathcal{G}|, \quad \text{s.t.} \sum_{y \in \mathcal{G}} p_y \geq \eta, \quad (4)$$

where p_y denotes the class prior probability of class y. The tail class group $\mathcal{G}_t$ is then defined as $\mathcal{Y} \setminus \mathcal{G}_h$. For illustration, Figure 1a shows the class distribution of a long-tailed CIFAR100, where classes are sorted by empirical frequency, and the head group corresponds to the smallest subset whose cumulative frequency exceeds η.

STANDARD Standard CP methods applied to long-tailed settings often exhibit substantial disparity between head and tail coverage. To quantify this effect, we evaluate APS-based Split CP on CIFAR100-LT under various target coverages $1 - \alpha$. The details of experiments can be found in Appendix A. The results in Figure 1b show that head classes are consistently over-covered, with conditional coverage lying above the target, while tail classes are significantly under-covered, leading to a large head–tail gap. The large imbalance of the label distribution causes this phenomenon because standard CP does not differentiate between classes, leading to frequent exclusion of minority classes from prediction sets. This observation motivates our approach to enforce fairer coverage across head and tail classes.

Tail-Aware Conformal Prediction (TACP)

To explicitly mitigate this head–tail coverage imbalance and improve the robustness of conformal prediction under long-tailed settings, we propose the **T**ail-**A**ware **C**onformal **P**rediction (TACP) method. TACP adaptively adjusts the label-ranking penalty by leveraging head–tail partition information, ensuring that prediction sets remain informative while reducing systematic coverage gaps between head and tail classes. Formally, the definition of TACP is given by:

$$s_{\text{TACP}}(\boldsymbol{x}, y) = s(\boldsymbol{x}, y) + \lambda \mathbb{I}(y \in \mathcal{G}_h)(o_{\boldsymbol{x}}(y) - k_r)^+, \quad (5)$$

where $\lambda \in \mathbb{R}^+$ and $k_r \in \mathbb{N}$ are hyperparameters; $()^+$ denotes the ReLU function; $\mathbb{I}$ is the indicator function; and $o_{\boldsymbol{x}}(y) = |\{y' \in \mathcal{Y} : \hat{\pi}_{y'}(\boldsymbol{x}) \geq \hat{\pi}_y(\boldsymbol{x})\}|$ denotes the ranking of y based on the estimated class posterior probability $\hat{\pi}(\boldsymbol{x})$. Here, $s(\boldsymbol{x}, y)$ can be any non-conformity scores. For example, the TACP-LAC method with $\lambda = 1$ and $k_r = 2$ can be derived by setting $s(\boldsymbol{x}, y)$ as one minus estimated class posterior probability:

$$s_{\text{TACP-LAC}}(\boldsymbol{x}, y) = \underbrace{1 - \hat{\pi}_y(\boldsymbol{x})}_{\text{LAC Score}} + \underbrace{\mathbb{I}(y \in \mathcal{G}_h)(o_{\boldsymbol{x}}(y) - 2)^+}_{\text{Selective Rank Regularization}}. \tag{6}$$

The key insight behind TACP is to utilize the LT information and selectively penalize label rankings for head, thereby reducing head-conditional coverage. To maintain the target marginal coverage level, this has the effect of increasing the tail-conditional coverage, thus narrowing the head-tail coverage gap. The label prediction set is then constructed:

$$\mathcal{C}_{\text{TACP}}(\boldsymbol{x}_{n+1}) := \{y \in \mathcal{Y} : s_{\text{TACP}}(\boldsymbol{x}_{n+1}, y) \leq \hat{q}_\alpha\}, \tag{7}$$

where $\hat{q}_\alpha$ is the $\lceil (1-\alpha)(n+1) \rceil / n$ quantile of the calibration scores $\{s_{\text{TACP}}(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$. Moreover, our TACP method enjoys the standard marginal coverage property:

Theorem 1 (TACP Coverage Guarantee)**.** *If $\{(X_i, Y_i)\}_{i=1}^n$ are i.i.d, then for any new i.i.d. draw (X_{n+1}, Y_{n+1}):*

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}_{TACP}(X_{n+1})) \geq 1 - \alpha \tag{8}$$

for the conformal prediction set $\mathcal{C}_{TACP}$ constructed in Eq.(7). Moreover, if we assume additionally a uniform random variable u to ensure the scores $\{s_{TACP}(X_i, Y_i)\}_{i=1}^{n+1}$ are almost surely distinct, then

$$\begin{aligned} s_{TACP}(X_i, Y_i) =& s(X_i, Y_i) \\ &+ \lambda \mathbb{I}(Y_i \in \mathcal{G}_h)(o_{X_i}(Y_i) - k_r + u)^+, \end{aligned} \tag{9}$$

and the following upper bound holds

$$\mathbb{P}(Y_{n+1} \in \mathcal{C}_{TACP}(X_{n+1})) \leq (1-\alpha) + \frac{1}{n+1}. \tag{10}$$

We now present a theorem showing the effectiveness of TACP in reducing head–tail conditional coverage disparity.

Theorem 2 (Improved Coverage Gap)**.** *Let E_{xy} be the event that the calibration dataset is fixed as $\{(X_i, Y_i)\}_{i=1}^n = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$. Suppose $\mathcal{C}_{TACP}$ and $\mathcal{C}_{STD}$ are the prediction sets constructed by TACP and the STANDARD method, respectively, based on the same non-conformity score. Then there exists k_r, such that the group-conditional coverage gap between head and tail groups satisfies:*

$$\begin{aligned} &P(Y_{n+1} \in \mathcal{C}_{TACP}(X_{n+1}) \mid Y_{n+1} \in \mathcal{G}_h, E_{xy}) \\ &- P(Y_{n+1} \in \mathcal{C}_{TACP}(X_{n+1}) \mid Y_{n+1} \in \mathcal{G}_t, E_{xy}) \\ \leq &P(Y_{n+1} \in \mathcal{C}_{STD}(X_{n+1}) \mid Y_{n+1} \in \mathcal{G}_h, E_{xy}) \\ &- P(Y_{n+1} \in \mathcal{C}_{STD}(X_{n+1}) \mid Y_{n+1} \in \mathcal{G}_t, E_{xy}). \end{aligned} \tag{11}$$

By introducing the selective rank regularization term, the adjusted acceptance threshold increases the probability of correctly including tail classes, while the head-conditional coverage is controlled by properly choosing parameter λ and k_r, thereby narrowing the coverage gap compared with STANDARD method using the same non-conformity score. A complete proof is given in Appendix B.

5 Soft TACP

While TACP addresses coverage imbalance between head and tail groups, a more fine-grained goal is to balance coverage across individual classes in LT settings (Löfström et al. 2015). In this section, we extend TACP beyond the binary head–tail partition to promote coverage balance across all classes, i.e., moving towards class-conditional coverage rather than merely head–tail coverage.

In the original TACP formulation Eq. (5), the indicator term plays an crucial role in penalizing head classes, thereby reducing the head–tail coverage gap as guaranteed by Theorem 2. However, this indicator term may be a hindrance when we want to move beyond the head-tail setting: according to the definition of $\mathcal{G}_h$ in Eq. (4), it naturally entails the head-tail partition, which may not align well with the goal of class-conditional coverage balance.

To address this limitation, we propose removing the dependence on the head–tail partition $\mathcal{G}_h$ and replacing the binary indicator with a soft, class-aware weighting scheme. This allows penalties to adapt continuously to class prior probabilities rather than enforcing a rigid head–tail split, enabling more nuanced control over coverage across classes.

Soft TACP Building on this idea, we introduce the **s**oft **T**ail-**A**ware **C**onformal **P**rediction (sTACP) method, which generalizes TACP by continuously reweighting penalties according to the estimated class prior $\hat{p}(y) \in [0, 1]$. Instead of the original hard 0–1 indicator, sTACP employs a smooth weighting function that reflects both the long-tailed structure and class-specific information:

$$s_{\text{sTACP}}(\boldsymbol{x}, y) = s(\boldsymbol{x}, y) + \lambda \hat{p}(y)(o_{\boldsymbol{x}}(y) - k_r)^+, \tag{12}$$

where $\lambda \in \mathbb{R}^+$ and $k_r \in \mathbb{N}$ are hyperparameters. This enables the label ranking penalty to adapt more smoothly to the long-tailed distribution without dependency to the head-tail partition. The efficiency of this extension is experimentally validated in Section 7.

6 Head-Tail Experiments

We evaluate the head- and tail-conditional coverages of STANDARD, Partition-Wise, and TACP methods on several versions of long-tailed CIFAR100 and ImageNet datasets using four representative non-conformity scores.

Experimental Setup

Datasets and Models In all experiments, the training, calibration, and test splits follow a long-tailed distribution, ensuring that the conformal calibration process operates under highly imbalanced conditions. For CIFAR100 (Krizhevsky, Hinton et al. 2009), we construct long-tailed variants following Cao et al. (2019b) with imbalance factors $\mu \in \{50, 100\}$, where $\mu = \frac{\max_i n_i}{\min_j n_j}$ and n_k denotes the number of samples from class k. For ImageNet (Deng et al. 2009), we create long-tailed subsets by sampling from a Pareto distribution with power parameter $\rho \in \{0.3, 0.6\}$ following Liu et al. (2019). All methods are evaluated using pretrained models from Li et al. (2021). Table 1 summarizes the datasets and models used in our experiments.

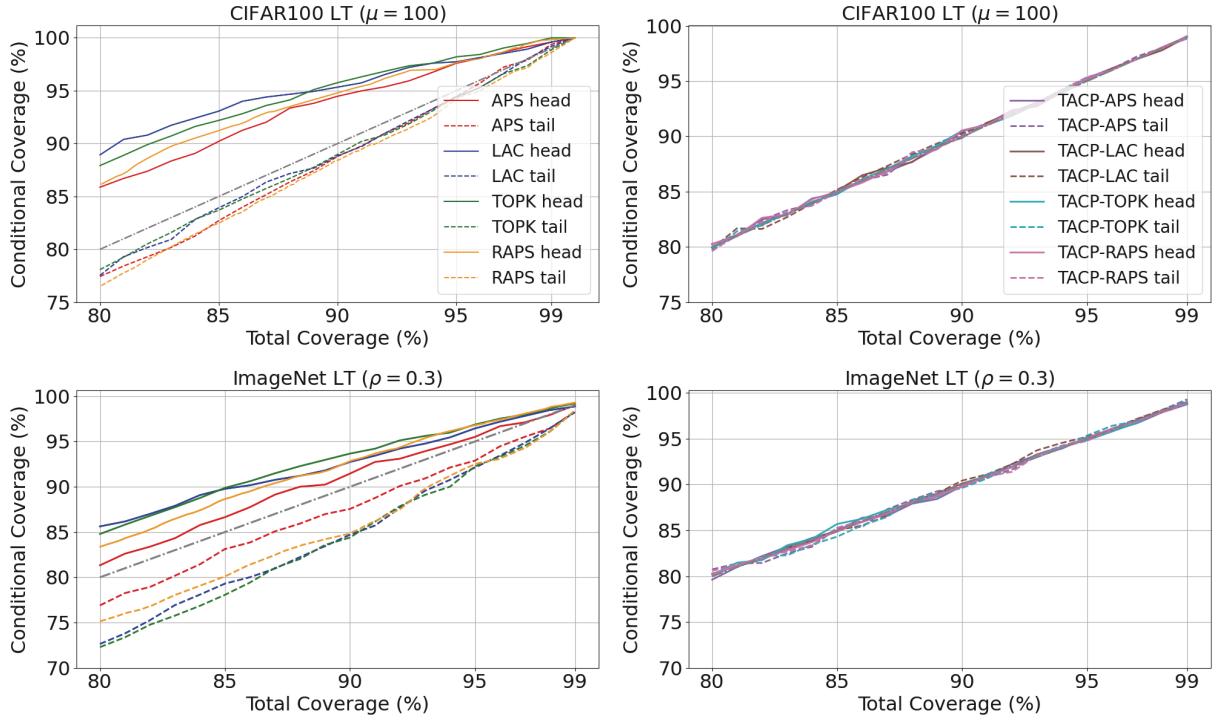


Figure 2: Head- and tail-conditional coverage versus total coverage for STANDARD (left column) and TACP (right column) across CIFAR100-LT ($\mu = 100$) and ImageNet-LT ($\rho = 0.3$). Results are shown for four non-conformity scores at $\eta = 50\%$. The gray dashed line indicates the ideal coverage (diagonal).

Statistics	CIFAR100 LT		ImageNet LT	
#Label	100	100	1000	1000
Pareto Power (ρ)	—	—	0.3	0.6
Imbalance Rate (μ)	100	50	8.3	25
#Train	10626	12354	115846	115846
#Test	5478	6329	8442	2453
Prediction Acc	59.80%	56.14%	48.21%	50.75%

Table 1: Specification of datasets. We evaluate the pretrained models' performance on long-tailed test datasets.

Baselines We compare three methods: **STANDARD**, **Partition-Wise**, and **TACP**. The Partition-Wise approach aims to reduce the head–tail coverage gap by dividing the calibration set into head and tail subsets and applying conformal prediction independently within each group. Additional details on these methods, along with the hyperparameter tuning procedure for TACP, are provided in Appendix A.

Non-conformity Scores We consider four scores: **APS**, which approximates X-conditional coverage (Romano, Sesia, and Candès 2020); **RAPS**, a regularized variant of APS to generate smaller prediction set (Angelopoulos et al. 2021); **LAC**, defined as one minus the softmax output (Sadinle, Lei, and Wasserman 2019); **TOPK**, which yields uniformly prediction set for all test samples (Angelopoulos et al. 2021). Definitions of the scores and the hyperparameters used in our experiments are provided in Appendix A.

Evaluation Metric Assume $\mathcal{I}_h = \{i \in [N_{\text{test}}] : y'_i \in \mathcal{G}_h\}$ and $\mathcal{I}_t = \{i \in [N_{\text{test}}] : y'_i \in \mathcal{G}_t\}$ be the indices of test examples $\{(\boldsymbol{x}'_j, y'_j)\}_{j=1}^{N_{\text{test}}}$ with head and tail classes, respectively. We give the metrics of head-tail coverage gap and average set size as follows:

$$\text{Cov-head} = 100 \times \frac{\sum_{i \in \mathcal{I}_h} \mathbb{I}(y'_i \in \mathcal{C}(\boldsymbol{x}'_i))}{|\mathcal{I}_h|}, \tag{13}$$

$$\text{Cov-tail} = 100 \times \frac{\sum_{i \in \mathcal{I}_t} \mathbb{I}(y'_i \in \mathcal{C}(\boldsymbol{x}'_i))}{|\mathcal{I}_t|}, \tag{14}$$

$$\text{CovGap-HT} = |\text{Cov-head} - \text{Cov-tail}|, \tag{15}$$

$$\text{AvgSize} = \frac{1}{N_{\text{test}}} \sum_{i=1}^{N_{\text{test}}} |\mathcal{C}(\boldsymbol{x}_i)|. \tag{16}$$

Results

We start with a brief summary of the experiments. In Experiment 1 (**Exp.1**), we compare the Cov-head and Cov-tail of STANDARD and TACP methods using four non-conformity scores on CIFAR100 LT and ImageNet LT datasets, showing that TACP can significantly reduce the CovGap-HT. In Experiment 2 (**Exp.2**), we evaluate the AvgSize and CovGap-

Score	Method	$\rho = 0.6$		$\rho = 0.3$		$\mu = 100$		$\mu = 50$	
		CovGap-HT	AvgSize	CovGap-HT	AvgSize	CovGap-HT	AvgSize	CovGap-HT	AvgSize
APS	STA	2.18±1.21	36.43±2.21	2.79±0.85	35.29±2.29	4.81±0.89	9.93±0.45	3.87±0.81	8.58±0.32
	PW	2.01±1.29	38.10±2.65	1.43±1.20	38.35±2.63	1.38±1.19	10.28±0.43	1.22±0.84	8.79±0.25
	TACP	**1.11±0.71**	33.98±2.40	**0.76±0.60**	34.07±2.22	**0.78±0.64**	10.25±0.46	**0.76±0.59**	8.55±0.26
LAC	STA	7.45±1.33	15.36±0.96	7.50±0.73	14.81±1.00	6.35±1.01	7.11±0.29	4.88±0.91	6.16±0.26
	PW	1.96±1.28	17.22±1.30	1.25±1.25	17.72±1.20	1.65±1.18	7.44±0.35	1.56±1.13	6.36±0.25
	TACP	**1.18±0.94**	17.92±1.40	**0.81±0.67**	19.16±1.91	**0.77±0.59**	7.86±0.36	**0.66±0.52**	6.47±0.25
TOPK	STA	8.83±1.69	23.15±1.45	9.45±0.89	20.77±1.22	6.47±0.84	10.90±0.50	5.51±0.82	8.58±0.36
	PW	1.87±1.53	24.06±2.46	1.31±0.98	24.25±1.98	1.33±1.16	11.01±0.47	1.27±0.97	8.71±0.40
	TACP	**1.02±0.84**	25.63±2.44	**0.58±0.48**	26.26±2.24	**0.82±0.61**	11.88±0.57	**0.77±0.57**	9.38±0.39
RAPS	STA	7.76±1.25	16.49±1.30	7.34±0.71	15.98±1.04	6.34±0.80	7.67±0.52	4.58±0.75	6.62±0.18
	PW	1.91±1.43	20.32±1.56	1.27±1.00	20.57±1.48	1.49±1.16	8.93±0.50	1.24±0.88	6.48±0.23
	TACP	**1.10±0.87**	20.67±1.58	**0.61±0.41**	21.76±2.03	**0.82±0.70**	9.62±0.68	**0.80±0.58**	6.76±0.24

Table 2: We report the average results±standard deviation over 100 different trails at $\alpha = 10\%$ and $\eta = 50\%$. Bold indicates the optimal CovGap-HT among all approaches for each non-conformity score.

HT of STANDARD, Partition-Wise, and TACP methods on several versions of CIAFR100 LT and ImageNet LT.

Exp.1: Cov-head and Cov-tail To evaluate the coverage disparity between head and tail classes, we compare the STANDARD and TACP methods on CIFAR100-LT ($\mu = 100$) and ImageNet-LT ($\rho = 0.3$) at a fixed partition $\eta = 50\%$, using four non-conformity scores. Specifically, we examine the results of Cov-head and Cov-tail.

Figure 2 illustrates the behavior of different methods as the total coverage level $1 - \alpha$ varies from 80% to 99% in steps of 1%. We observe that under LT label distributions, the STANDARD method suffers from a coverage bias between head and tail classes, with the Cov-head (solid curves) for all scores consistently lying above the diagonal and the Cov-tail (dashed curves) remaining below it. In contrast, the TACP method effectively narrows the gap between the Cov-head and the Cov-tail, as shown by the almost overlapping dashed and solid lines. The results show that, compared with STANDAND, TACP can significantly alleviate the head-tail coverage imbalance and mitigate the under coverage of tail classes across all scores and datasets in LT settings.

Exp.2: CovGap-HT and AvgSize We compare the STANDARD (STA), Partition-Wise (PW), and TACP methods in terms of CovGap-HT and AvgSize across multiple LT datasets: CIFAR100-LT with $\mu \in \{50, 100\}$ and ImageNet-LT with $\rho \in \{0.3, 0.6\}$, evaluated at a miscoverage level $\alpha = 10\%$ and head–tail partition $\eta = 50\%$. The results are summarized in Table 2.

We can observe that TACP consistently achieves a substantial reduction in CovGap-HT without compromising prediction set informativeness compared to the STANDARD method. For example, applying TOPK as base score on ImageNet LT ($\rho = 0.6$), the CovGap-HT of the STANDARD-TOPK method is 8.83%, whereas TACP-TOPK reduces it to 1.02% with only a slight increase in AvgSize from 23.15 to 25.63. Overall, these results demonstrate the effectiveness of TACP in narrowing CovGap-HT while maintaining efficient prediction sets. Additional results over a broader range of η (Appendix A) further confirm the robustness of TACP.

7 Class-conditional Experiments

In this section, we compare the class-conditional results of STANDARD, CLASSWISE, CLUSTER, RC3P, and sTACP on ImageNet-LT across four non-conformity scores.

Experimental Setup

Baselines We use the same datasets and pretrained models described in Section 6. We evaluated five methods: STANDARD, CLASSWISE, CLUSTER, RC3P, and sTACP. Details of these baselines and hyperparameters for finetuning of sTACP are provided in Appendix C.

Evaluation Metric We evaluate the performance on three metrics: **coverage**, which measures the empirical marginal coverage; **set size** (AvgSize), which reflects the efficiency of the prediction set; **class-conditional coverage gap** (CovGap), which quantifies coverage disparities across classes. Details are provided in Appendix C.

Results

We provide a brief overview of the experimental setup. Experiment 3 (**Exp.3**) reports CLASSWISE results on ImageNet-LT using four different scores. Experiment 4 (**Exp.4**) compares STANDARD, CLUSTER, RC3P, and sTACP across two long-tailed variants of ImageNet-LT. Experiment 5 (**Exp.5**) examines performance on ImageNet-LT ($\rho = 0.6$) under a different target coverage, with $\alpha = 5\%$.

Exp.3: CLASSWISE Method To examine the limitations of the CLASSWISE (CW) method under LT settings, we conduct an experiment on ImageNet-LT ($\rho = 0.6$) at $\alpha = 10\%$ using four different non-conformity scores. As shown

Score	Method	$\rho = 0.6$			$\rho = 0.3$		
		CovGap	Coverage	AvgSize	CovGap	Coverage	AvgSize
APS	STA	19.00±0.00	89.78±0.01	35.60±2.27	13.37±0.35	90.07±0.61	38.39±1.54
	CLUS	17.84±0.00	89.66±0.01	37.54±1.43	13.41±0.00	90.62±0.00	42.80±1.28
	RC3P	18.99±0.61	89.31±0.65	30.20±1.88	14.21±0.31	89.45±0.45	43.50±2.22
	sTACP	**15.86±0.81**	89.53±1.12	36.52±3.66	**12.76±0.31**	90.04±0.61	35.48±2.08
LAC	STA	18.87±0.01	90.20±0.01	15.30±0.81	14.35±0.36	89.90±0.63	16.94±0.67
	CLUS	17.80±0.00	89.10±0.00	16.19±0.05	14.05±0.00	90.82±0.00	19.59±0.10
	RC3P	19.17±0.59	89.19±0.70	30.45±2.09	14.24±0.32	89.55±0.45	43.10±2.00
	sTACP	**15.98**±0.78	89.54±1.07	37.57±3.40	**12.66±0.28**	90.09±0.59	41.72±1.72
TOPK	STA	18.87±0.01	90.12±0.01	23.28±1.23	14.38±0.41	89.96±0.61	23.55±1.02
	CLUS	17.59±0.00	89.57±0.00	24.24±0.15	13.89±0.00	91.61±0.00	27.96±0.20
	RC3P	18.99±0.61	89.39±0.67	31.01±2.15	14.21±0.31	89.67±0.47	45.62±2.32
	sTACP	**15.83±0.77**	89.55±1.08	38.99±3.54	**12.80±0.28**	90.05±0.58	37.38±1.71
RAPS	STA	18.45±0.93	89.98±1.07	16.49±1.30	14.25±0.33	89.98±0.60	18.32±1.04
	CLUS	17.69±0.00	90.13±0.00	18.53±0.09	13.92±0.00	90.99±0.00	23.54±1.25
	RC3P	18.99±0.61	89.35±0.66	30.92±2.14	14.19±0.31	89.70±0.46	45.56±2.31
	sTACP	**16.04±0.82**	89.54±1.10	33.68±3.45	**12.90±0.31**	90.04±0.61	32.85±1.65

Table 3: Performance on ImageNet LT at $\alpha = 10\%$. We report the average results±standard deviation for 100 different trials. Bold indicates the optimal CovGap for each non-conformity score.

Score	CovGap	Coverage	AvgSize
APS	10.00±0.02	99.46±0.31	944.48±2.42
LAC	9.98±0.01	99.51±0.22	942.51±1.43
TOPK	9.98±0.02	99.56±0.27	943.16±1.23
RAPS	10.00±0.02	99.54±0.33	943.15±1.22

Table 4: Performance of CLASSWISE method on ImageNet LT ($\rho = 0.6$) at $\alpha = 10\%$.

Method	CovGap	Coverage	AvgSize
STA	9.62±0.01	94.88±0.01	67.49±5.62
CW	5.00±0.01	99.90±0.16	983.53±2.86
CLUS	———	———	———
RC3P	15.28±0.69	89.57±0.66	31.80±2.28
sTACP	8.08±0.00	95.30±0.00	76.83±0.17

Table 5: Performance with base score APS on ImageNet LT ($\rho = 0.6$) at $\alpha = 5\%$.

in Table 4, the CW approach frequently outputs the entire label set $\mathcal{Y}$, leading to a trivial average coverage of 100%. Such excessively large prediction sets are uninformative and of limited value for downstream tasks.

Exp.4: sTACP Method We compare STANDARD (STA), CLUSTER (CLUS), RC3P, and sTACP methods on two versions of ImageNet LT $\rho \in \{0.3, 0.6\}$ at $\alpha = 10\%$. As shown in Table 3, sTACP consistently reduces CovGap with only a modest increase in AvgSize compared with other baselines. For instance, the CovGap of STA and sTACP decreases from 19.00% to 15.86% when applying APS as the base score on ImageNet LT ($\rho = 0.3$), whereas their AvgSize remains similar. In summary, sTACP method consistently outperforms the baselines under LT distributions, achieving smaller class-conditional coverage gaps with the same score.

Exp.5: Evaluation at Stricter Coverage To further evaluate robustness under stricter coverage requirements, we apply the APS score to all baselines at $\alpha = 5\%$ on ImageNet-LT ($\rho = 0.6$). Table 5 shows that CW generates uninformative prediction sets with an average size of 984. RC3P fails in this setting, likely due to the instability of class-specific top-k error estimates under severe data imbalance. For CLUSTER, the combination of long-tailed distributions and a high coverage requirement causes the clustering to collapse into a single "null" cluster, effectively reducing the method to STANDARD and rendering its results uninformative.

In contrast, sTACP achieves substantially smaller CovGap while maintaining efficient prediction sets. Additional results using the LAC at $\alpha = 5\%$ are reported in Appendix C.

8 Conclusion

In this paper, we propose TACP to achieve a more balanced performance across head and tail labels under long-tailed distributions. By selectively penalizing label ranks for head labels, TACP reduces the prediction set size for head labels, which indirectly improves the tail-conditional coverage, thus narrowing the conditional coverage gap. We also provide a theoretical analysis demonstrating TACP's effectiveness in reducing this gap. We further proposed sTACP method to reduce the class-conditional coverage gap in long-tailed settings. Experimental results demonstrate that our proposed methods consistently improve head-tail and class-conditional coverage gap across several benchmarks.

Acknowledgments

This work was supported by the National Research Foundation, Singapore, under its RSS Scheme (NRF-RSS2022-009), and the Singapore Global AI VP Ruqola Project.

References

Angelopoulos, A. N.; Bates, S.; Jordan, M. I.; and Malik, J. 2021. Uncertainty Sets for Image Classifiers using Conformal Prediction. In *ICLR*.

Barber, R. F.; Candes, E. J.; Ramdas, A.; and Tibshirani, R. J. 2021. Predictive inference with the jackknife+. *Ann. Stat.*, 49(1): 486–507.

Bates, S.; Candès, E.; Lei, L.; Romano, Y.; and Sesia, M. 2023. Testing for outliers with conformal p-values. *Ann. Stat.*, 51(1): 149–178.

Brodersen, K. H.; Ong, C. S.; Stephan, K. E.; and Buhmann, J. M. 2010. The balanced accuracy and its posterior distribution. In *ICPR*.

Buda, M.; Maki, A.; and Mazurowski, M. A. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Netw.*, 106: 249–259.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019a. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*.

Cao, K.; Wei, C.; Gaidon, A.; Aréchiga, N.; and Ma, T. 2019b. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *NeurIPS*.

Caprio, M. 2025. The joys of categorical conformal prediction. *arXiv preprint arXiv:2507.04441*.

Caprio, M.; Stutz, D.; Li, S.; and Doucet, A. 2025. Conformalized Credal Regions for Classification with Ambiguous Ground Truth. *Trans. Mach. Learn. Res.*

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, 16: 321–357.

Cui, Y.; Jia, M.; Lin, T.; Song, Y.; and Belongie, S. J. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *CVPR*.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*.

Ding, T.; Angelopoulos, A.; Bates, S.; Jordan, M. I.; and Tibshirani, R. J. 2023. Class-Conditional Conformal Prediction with Many Classes. In *NeurIPS*.

Grigorescu, S.; Trasnea, B.; Cocias, T.; and Macesanu, G. 2020. A survey of deep learning techniques for autonomous driving. *J. Field Robot.*, 37(3): 362–386.

Guan, L.; and Tibshirani, R. 2022. Prediction and outlier detection in classification problems. *J. R. Stat. Soc. Ser. B-Stat. Methodol.*, 84(2): 524–546.

Huang, J.; Xi, H.; Zhang, L.; Yao, H.; Qiu, Y.; and Wei, H. 2024. Conformal Prediction for Deep Classifier via Label Ranking. In *ICML*.

Kasa, K.; and Taylor, G. W. 2023. Empirically validating conformal prediction on modern vision architectures under distribution shift and long-tailed data. *arXiv preprint arXiv:2307.01088*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lee, K.; and Zhang, Y. 2025. Leave-One-Out Stable Conformal Prediction. In *ICLR*.

Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R. J.; and Wasserman, L. 2018. Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.*, 113(523): 1094–1111.

Lei, J.; Robins, J.; and Wasserman, L. 2013. Distribution-free prediction sets. *J. Am. Stat. Assoc.*, 108(501): 278–287.

Li, S.; Gong, K.; Liu, C. H.; Wang, Y.; Qiao, F.; and Cheng, X. 2021. MetaSAug: Meta Semantic Augmentation for Long-Tailed Visual Recognition. In *CVPR*.

Liu, K.; Zeng, H.; Huang, J.; Zhuang, H.; Vong, C.-M.; and Wei, H. 2024. C-Adapter: Adapting Deep Classifiers for Efficient Conformal Prediction Sets. *arXiv preprint arXiv:2410.09408*.

Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; and Yu, S. X. 2019. Large-Scale Long-Tailed Recognition in an Open World. In *CVPR*.

Löfström, T.; Boström, H.; Linusson, H.; and Johansson, U. 2015. Bias reduction through conditional conformal prediction. *Intell. Data Anal.*, 19(6): 1355–1375.

Lu, C.; Lemay, A.; Chang, K.; Höbel, K.; and Kalpathy-Cramer, J. 2022. Fair conformal predictors for applications in medical imaging. In *AAAI*.

Papadopoulos, H.; Proedrou, K.; Vovk, V.; and Gammerman, A. 2002. Inductive confidence machines for regression. In *ECML*.

Plassier, V.; Fishkov, A.; Guizani, M.; Panov, M.; and Moulines, E. 2025. Probabilistic conformal prediction with approximate conditional validity. In *ICLR*.

Podkopaev, A.; and Ramdas, A. 2021. Distribution-free uncertainty quantification for classification under label shift. In *UAI*.

Romano, Y.; Patterson, E.; and Candes, E. 2019. Conformalized quantile regression. In *NeurIPS*.

Romano, Y.; Sesia, M.; and Candès, E. J. 2020. Classification with Valid and Adaptive Coverage. In *NeurIPS*.

Sadinle, M.; Lei, J.; and Wasserman, L. 2019. Least ambiguous set-valued classifiers with bounded error levels. *J. Am. Stat. Assoc.*, 114(525): 223–234.

Seedat, N.; Jeffares, A.; Imrie, F.; and van der Schaar, M. 2023. Improving adaptive conformal prediction using self-supervised learning. In *AISTATS*.

Sesia, M.; and Romano, Y. 2021. Conformal Prediction using Conditional Histograms. In *NeurIPS*.

Shi, F.; Ong, C. S.; and Leckie, C. 2013. Applications of class-conditional conformal predictor in multi-class classification. In *ICMLA*.

Shi, Y.; Ghosh, S.; Belkhouja, T.; Doppa, J.; and Yan, Y. 2024. Conformal Prediction for Class-wise Coverage via Augmented Label Rank Calibration. In *NeurIPS*.

Tan, J.; Wang, C.; Li, B.; Li, Q.; Ouyang, W.; Yin, C.; and Yan, J. 2020. Equalization loss for long-tailed object recognition. In *CVPR*.

Tibshirani, R. J.; Foygel Barber, R.; Candes, E.; and Ramdas, A. 2019. Conformal prediction under covariate shift. In *NeurIPS*.

Vovk, V. 2013. Conditional validity of inductive conformal predictors. *Mach. Learn.*, 92(2-3): 349–376.

Vovk, V. 2015. Cross-conformal predictors. *Ann. Math. Artif. Intell.*, 74(1): 9–28.

Vovk, V.; Gammerman, A.; and Shafer, G. 2005. *Algorithmic learning in a random world*. Springer.

Xi, H.; Huang, J.; Liu, K.; Feng, L.; and Wei, H. 2025. Does confidence calibration improve conformal prediction? *Trans. Mach. Learn. Res.*

Xu, R.; Chen, C.; Sun, Y.; Venkitasubramaniam, P.; and Xie, S. 2025. Wasserstein-regularized conformal prediction under general distribution shift. In *ICLR*.

Zhou, Y.; and Sesia, M. 2024. Conformal Classification with Equalized Coverage for Adaptively Selected Groups. In *NeurIPS*.