CONVERTING EFFICIENCY TO CLASSWISE COVERAGE: CLUSTER-FREQUENCY CONFORMAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Conformal prediction guarantees marginal coverage under exchangeability, but coverage can vary substantially across classes. We propose Cluster-Frequency Conformal Prediction (CFCP), which forms probability vectors from softly pooled cluster-label frequencies, prior smoothing, and a reliability-weighted fallback before standard conformal scoring. Independent final calibration preserves marginal validity. Across seven image and text evaluation settings, including three with long-tailed training and four score families, we examine whether efficient prediction sets provide headroom for classwise coverage. CFCP can achieve smaller prediction sets at matched marginal coverage, motivating comparisons of classwise coverage at equal size budgets. At CFCP's nominal mean set size, paired comparisons favor CFCP in 22 of 28 cells and the selected best baseline in one, with five inconclusive and gains up to 22.47 percentage points. Comparisons averaged over shared size ranges favor CFCP in 24 cells and the baseline in three, with one inconclusive. These results support an empirical classwise coverage-size advantage without establishing classwise coverage guarantees.

1 INTRODUCTION

Conformal prediction (CP) provides finite-sample marginal coverage guarantees for set-valued predictions under exchangeability (Vovk et al., 2005; Angelopoulos & Bates, 2021). Marginal validity can nevertheless coexist with undercoverage of particular classes or input-defined subpopulations. This motivates methods that adapt uncertainty estimates to heterogeneity in the data, especially in many-class settings where relatively few calibration examples are available per class.



Figure 1: Clusters by learned image representations. Samples from two clusters of the CIFAR10 data. Left: samples from a highly homogeneous cluster (all labeled horse). Right: samples from a label-heterogeneous cluster.

Several existing approaches address this heterogeneity through finer calibration: per class, per label cluster, or per neighborhood. We investigate a complementary direction: whether local label-composition information in learned representations can improve the probability vector supplied to a conformal score. Our hypothesis is that learned embeddings often organize examples into subpopulations with different label compositions (Figure 1). Local empirical frequencies may therefore provide information about label ambiguity that is not fully captured by the base model's probability vector.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We propose Cluster-Frequency Conformal Prediction (CFCP), which uses clustering to obtain a compact, regularized estimate of local label composition and refine the probability vector supplied to standard conformal prediction. CFCP separates this probability estimation step from final calibration. Its formal guarantee concerns marginal coverage, while improvements in classwise coverage and efficiency remain empirical objectives.

Our empirical thesis is that set-size efficiency provides headroom to improve classwise coverage within a size budget. Coverage-matched comparisons provide an initial diagnostic, but smaller sets can accompany lower Class Coverage (COV). We therefore evaluate the coverage-size tradeoff at a fixed nominal mean set size and over shared size ranges. Throughout, the classwise evaluation target remains 0.9 as the requested calibration level varies.

CFCP and Neighborhood Conformal Prediction (NCP) lead the reported COV comparisons in most settings. This pattern supports the empirical value of representation-space locality and motivates a closer comparison of how the two methods incorporate local information into conformal prediction.

Our contributions are:

- We introduce a cluster-frequency estimator that forms query-specific probability vectors from softly pooled cluster-label frequencies, prior smoothing and a reliability-weighted fallback to the base model's probabilities. The resulting vectors can be used with standard conformal scores.
- We establish that, under exchangeability and independent final calibration, this probability substitution preserves the standard finite-sample marginal coverage guarantee, regardless of the accuracy of the estimated probabilities. The guarantee also holds under the implemented nonempty-set convention.
- We conduct a comprehensive empirical evaluation across seven settings, including three with long-tailed training, using paired COV comparisons at the nominal size budget and over shared prediction-set size ranges.
- We analyze the computational cost of the compact cluster representation and compare its prediction-time performance with the evaluated implementation of the Neighborhood Conformal Prediction baseline.

1.1 Background and Related Work

1.1.1 Formal Definitions and Notations

Let a predictor and a nonconformity score s be fitted independently of a final calibration sample. For n calibration scores $s_i = s(X_i, Y_i)$, split conformal prediction (Split) uses the threshold

$$\widehat{t}_\alpha = s_{(k_\alpha)}, \qquad k_\alpha = \lceil (n+1)(1-\alpha) \rceil ,$$

where $s_{(j)}$ is the j-th order statistic and $s_{(n+1)} = +\infty$ by convention. The prediction set is $\widehat{\Gamma}_\alpha(x) = \{y : s(x,y) \le \widehat{t}_\alpha\}$. Then, the marginal coverage is $\mathbb{P}\{Y \in \widehat{\Gamma}_\alpha(x)\}$. Conditional exchangeability of the calibration and test examples gives marginal coverage at least $1-\alpha$. Randomized scores use independent auxiliary randomization with the same law for calibration and prediction.

Least Ambiguous Classifier (LAC) applies Split to the softmax score $1 - \hat{p}_x(y)$. It returns the set of labels whose predicted probability is at least the calibrated probability threshold $1 - \widehat{t}_\alpha$. Adaptive set constructions such as APS, regularized APS (RAPS) and Sorted Adaptive Prediction Sets (SAPS) use probability and rank information to adapt prediction-set size (Romano et al., 2020; Angelopoulos et al., 2020; Huang et al., 2023). These methods are strong baselines, but they primarily target marginal validity and set size, not disparities in coverage across classes or subpopulations.

To address this, class-conditional and many-class conformal methods calibrate more finely across labels. Mondrian conformal prediction and Conditional Inductive Conformal Prediction (ICP) target label-conditional behavior (Vovk, 2012), while Ding et al. (2023) stabilize many-class classwise calibration by clustering labels with similar conformity-score distributions (CCP). Ding et al. (2025) further study classwise reliability in long-tailed classification by reweighting predicted probabilities of under-represented classes. Shi et al. (2024) follows CCP while improving efficiency by classwise label ranking (RC3P).

A complementary direction localizes conformal prediction in feature space. Localized CP (LCP) prioritizes local neighborhood calibration points Guan (2023). Neighborhood CP (NCP) develops LCP and uses nearby examples in a learned embedding space (Ghosh et al., 2023). Posterior CP (PCP) models conformity scores through latent subgroup mixtures (Zhang & Candès, 2024), and recent work on local weighting and subgroup discovery further highlights the value of covariate-aware uncertainty quantification (Hore & Barber, 2023; Gil et al., 2024).

CFCP also connects to the classical Venn prediction framework, which constructs probability forecasts from empirical label frequencies within categories defined by a permutation-equivariant taxonomy. This provides a historical precedent for using local label composition to modify probability estimates. Venn-Abers predictors instantiate this principle through categories induced by isotonic regression on classifier scores Vovk et al. (2003); Vovk & Petej (2012).

CFCP adapts this reference-class perspective through soft pooling of label frequencies in learned representation space, with smoothing and a reliability-weighted fallback to the base predictor. This connection motivates introducing locality into the probability vector before conformal calibration. The theoretical distinction is essential: Venn predictors consider each candidate test label and return multiple probability forecasts, with validity attached to that construction. CFCP instead produces a single fitted vector q_x, for which we do not establish a Venn-style probability-calibration guarantee. Its finite-sample marginal coverage follows from the separate, protected split-conformal calibration step.

Within localized conformal prediction, NCP provides a direct comparison. CFCP and NCP differ along two axes. First, they apply local information at different stages: NCP weights neighboring calibration scores to obtain a query-specific conformal threshold, whereas CFCP replaces π_x by q_x before standard scoring and global split-conformal calibration. Second, they summarize locality differently: NCP uses distance-weighted calibration neighbors, CFCP pools label frequencies through soft assignments to learned centroids, smooths cluster distributions toward a global class prior and uses a reliability-weighted fallback. The comparison therefore evaluates these complete localization strategies. Inspired by Venn's use of category-level label frequencies, CFCP refines class probabilities and rankings before conformal calibration, offering a potential advantage over NCP's threshold localization when local label composition reveals class structure that the base predictor fails to capture.

CFCP and NCP also differ in their storage and computational requirements. Appendix C.2 provides a detailed comparison.

2 METHODS

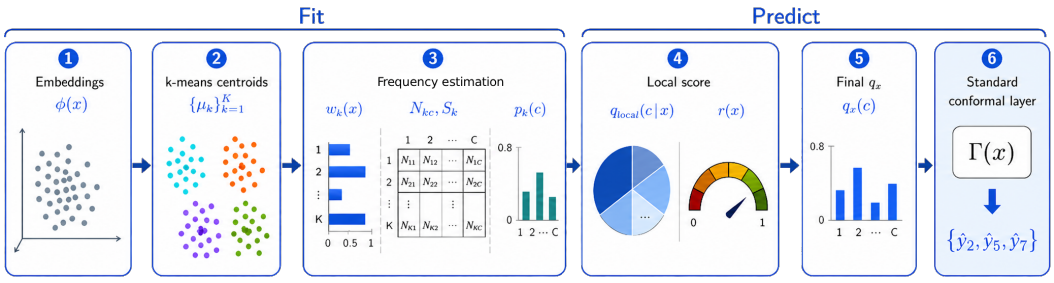


Figure 2: CFCP Pipeline Illustration. Learned representations (1) are clustered (2) and frequency-sample labels are accumulated as soft cluster counts (3). Smoothed cluster distributions are mixed according to the test point's soft assignment weights and reliability-shrunk to form probabilities (4), which are passed to a standard conformal set constructor (5+6). See full notation list in Table 4 in the appendix.

CFCP implements locality in the probability vector, q_x, which combines soft cluster-label frequencies with prior smoothing and a reliability-weighted fallback (Figure 2). Cluster geometry is fitted on D_{train}, and label frequencies on a designated sample D_{freq}, the frequency source depends on the experimental protocol (Appendix C.1.2), the complete estimator is fixed before final calibration.

Standard scores applied to q_x then use the untouched final calibration pool D_q to obtain a global threshold. Appendix A provides full definitions.

2.1 LOCAL CLUSTER-FREQUENCY MODEL

Let f be a fitted classifier over C classes and $\phi(x) \in \mathbb{R}^d$ its representation. We normalize embeddings and fit spherical k-means on the training representations, obtaining centroids $\mu_1, \ldots, \mu_K$. For a query x, let $\kappa_j(x)$ be its j-th nearest centroid, $j = 1, \ldots, m$, and let $\delta_j(x)$ be the corresponding cosine similarity. Define

$$w_j(x) = \frac{\exp(\delta_j(x)/\tau)}{\sum_{\ell=1}^{m} \exp(\delta_\ell(x)/\tau)}, \qquad a_k(x) = \sum_{j=1}^{m} w_j(x)\mathbf{1}\{\kappa_j(x) = k\}. \tag{1}$$

Here $\tau > 0$ controls assignment sharpness, and $a_k(x) = 0$ outside the retrieved centroids.

Let D_{freq} be the designated labeled frequency sample, disjoint from the final calibration sample D_q. With $a_{ik} = a_k(x_i)$, the soft cluster-label counts and supports are

$$N_{kc} = \sum_{(x_i,y_i)\in D_{\text{freq}}} a_{ik}\mathbf{1}\{y_i = c\}, \qquad S_k = \sum_{(x_i,y_i)\in D_{\text{freq}}} a_{ik}. \tag{2}$$

Each example distributes unit total weight over its neighboring centroids. A global prior $\pi \in \Delta^{C-1}$ and smoothing strength $\beta > 0$ give

$$\widehat{p}_k(c) = \frac{N_{kc} + \beta\pi_c}{S_k + \beta}, \qquad q_x^{\text{local}}(c) = \sum_{j=1}^{m} w_j(x)\widehat{p}_{\kappa(x)}(c). \tag{3}$$

The final vector interpolates this local summary with the base model's sample-specific probability vector π_x:

$$q_x(c) = r(x)q_x^{\text{local}}(c) + (1 - r(x))\pi_x(c). \tag{4}$$

We use

$$\text{conf}(x) = \left(\max_j w_j(x)\right)^{\gamma}, \qquad \text{sup}(x) = \sum_j w_j(x)s_j, \qquad r(x) = \text{conf}(x)\frac{\text{sup}(x)}{\text{sup}(x) + \beta_{\text{sup}}},$$

with $\gamma > 0$ and $\beta_{\text{sup}} > 0$. This rule reduces the local component's weight when assignments are diffuse or support is low. The weight uses geometry and support as heuristic measures of reliability. Its empirical contribution is examined in Section 4.1. At $r(x) = 0$, $q_x = \pi_x$ exactly.

K and m are tuned on development data. Appendix A specifies the geometry- and support-based defaults for τ, β, and β_{sup}, and the default $\gamma = 1$. These defaults can be overridden by tuning τ, β, β_{sup}, and γ on the same development data. All fitted quantities and tuning choices are fixed before final calibration.

2.2 CONFORMAL PREDICTION

CFCP separates probability estimation from set construction. We compute calibration scores from q_x using LAC, APS, RAPS, or SAPS, and calibrate the corresponding scalar threshold on D_q. The same score is used at prediction time. The implementation enforces nonempty sets by adding the top-ranked label when the raw thresholded set is empty. This enlargement preserves the marginal coverage lower bound.

3 THEORETICAL ANALYSIS

We state two results with distinct roles. Theorem 1 is the standard finite-sample validity result applied to CFCP's fitted score. Theorem 2 recalls a classical population oracle characterization to motivate posterior-based LAC scoring. Neither result establishes that the fitted CFCP estimator improves local or classwise coverage, and we evaluate those outcomes empirically.

Setting. Let $\eta_c(x) = \Pr\{Y = c \mid X = x\}$. The complete fitted model $\widehat{M}$, including the representation, frequencies, priors, and tuning choices, fixes q_x and the score before final calibration on D_q, with $n_q = |D_q| \geq 1$. Fix $\alpha \in (0, 1)$ independently of this final sample. Write $\widehat{\Gamma}$ for the raw score-thresholded set and $\widehat{\Gamma}^+$ for its nonempty version. The calibration threshold is the k_α-th order statistic, where

$$k_\alpha = \lceil (n_q + 1)(1 - \alpha) \rceil, \qquad \nu_{n_q,\alpha} = \frac{k_\alpha}{n_q + 1},$$

with $s_{(n_q+1)} = +\infty$. Coverage probabilities condition on $\widehat{M}$ and average over final calibration, the test point, and auxiliary randomization. Appendix B provides the assumptions and proofs.

Theorem 1 (Finite-sample marginal validity)**.** *If calibration and test examples are exchangeable conditional on $\widehat{M}$, and their auxiliary randomization is independent with the same law, then, for each supported score constructor,*

$$\Pr\{Y_{n_q+1} \in \widehat{\Gamma}(X_{n_q+1}) \mid \widehat{M}\} \geq \nu_{n_q,\alpha} \geq 1 - \alpha.$$

The same lower bound holds for $\widehat{\Gamma}^+$.

This standard rank argument (Vovk et al., 2005) requires no accuracy assumption on q_x. Frequencies may be estimated from the model-training data when final calibration remains independent of fitting and tuning.

Classical oracle motivation. The next result is recalled from the least-ambiguous classifier characterization of Sadinle et al. (2019). We state it in LAC notation and include its standard proof for completeness.

Theorem 2 (Classical population oracle for LAC)**.** *Suppose $q_x = \eta(x)$ and $0 < \alpha < 1$. Choose $\tau_\alpha > 0$ and $\xi \in [0, 1]$ so that the randomized inclusion probabilities*

$$h_c^*(x) = \mathbf{1}\{\eta_c(x) > \tau_\alpha\} + \xi \mathbf{1}\{\eta_c(x) = \tau_\alpha\}$$

attain marginal coverage $1 - \alpha$. The corresponding population LAC predictor minimizes $\mathbb{E}|\Gamma(X)|$ over all measurable, possibly randomized predictors with marginal coverage at least $1 - \alpha$, where prediction randomization is independent of the unknown label given X.

Efficiency headroom and recalibration. For a fixed fitted score, calibration sample and score randomization, increasing requested coverage weakly increases the conformal threshold. The raw and implemented nonempty sets are nested, so classwise coverage at a fixed evaluation target cannot decrease. An initial set-size advantage therefore provides potential budget for expansion. Its realized classwise benefit is evaluated through the coverage-size frontier.

3.1 Computational complexity

For n_{train} embeddings of dimension d, full-batch spherical k-means with T iterations costs $O(T n_{\text{train}} K d)$. With a flat centroid index, similarity calculation and dense label-distribution mixing cost $O(Kd + mC)$ per query, apart from top-m selection and the chosen conformal score's usual cost. Centroids and cluster distributions require $O(Kd + KC)$ storage. Frequency accumulation costs $O(n_{\text{freq}} K d + n_{\text{freq}} m + KC)$, apart from the same retrieval-selection overhead. With K, m, C, and d fixed, these fitted-model prediction costs do not grow with the frequency-sample size. Appendix A.4 details retrieval and score-construction costs.

4 Experiments

We evaluate CFCP through extensive experiments, comparing it with strong baselines across a diverse set of classification tasks and model architectures. Code and the selected configurations for reproducing the experiments is available at `https://anonymous.4open.science/r/CFCP-C039/`.
Full details on model training and experimental setup are available in Appendix C.1.

Table 1: Within-family comparison of CFCP against baselines at common coverage-matched target of 90%. Values are means over five repeated dataset splits. **Bold** marks the best class coverage within each family and, for set size, the lower value between CFCP and the best-coverage baseline.

		CFCP (ours)		Split		ICP		CCP		NCP		RC3P	
Dataset	**Score**	COV↑	Size↓	COV↑	Size↓	COV↑	Size↓	COV↑	Size↓	COV↑	Size↓	COV↑	Size↓
CIFAR-100	LAC	**0.593**	**6.6**	0.516	8.0	0.568	11.2	0.534	9.7	0.582	8.9	0.568	11.2
	APS	0.610	**8.5**	0.516	11.1	0.560	14.6	0.530	12.2	**0.630**	13.7	0.582	11.3
	RAPS	**0.597**	**7.8**	0.556	8.3	0.570	11.6	0.580	10.6	0.562	8.3	0.564	11.6
	SAPS	**0.588**	**7.5**	0.558	8.2	0.582	11.5	0.568	10.3	0.576	7.8	0.548	10.9
ImageNet	LAC	**0.604**	**2.6**	**0.604**	2.7	0.493	6.8	0.577	4.4	**0.604**	2.7	0.588	11.1
	APS	0.735	**13.3**	**0.758**	55.0	0.708	75.6	0.743	61.0	0.748	30.1	0.723	24.3
	RAPS	0.644	**3.8**	0.641	3.8	**0.699**	23.4	0.626	16.6	0.644	3.7	0.575	11.2
	SAPS	**0.616**	3.1	0.611	**3.0**	0.499	10.2	0.597	17.3	0.610	3.0	0.581	11.2
ImageNet-V2	LAC	0.512	**3.1**	0.508	3.2	**0.577**	22.2	0.496	6.4	0.510	3.2	0.482	11.1
	APS	0.728	**21.6**	0.737	71.1	**0.775**	120.1	0.731	78.5	0.729	48.2	0.692	47.4
	RAPS	0.568	**4.6**	0.562	4.6	0.518	11.4	0.551	14.9	0.565	4.5	**0.611**	24.2
	SAPS	**0.515**	**3.4**	0.512	**3.4**	0.475	10.6	0.494	15.6	0.512	3.4	0.488	11.2
WOS-46985	LAC	**0.654**	**1.3**	0.609	1.2	0.619	4.2	0.591	5.2	0.652	1.3	0.629	4.0
	APS	0.667	**1.7**	0.552	3.4	0.610	6.1	0.565	6.9	**0.678**	2.0	0.639	4.3
	RAPS	0.667	**1.7**	0.560	2.7	0.613	6.0	0.570	6.2	**0.670**	1.9	0.647	4.4
	SAPS	**0.719**	1.9	0.591	1.3	0.613	5.8	0.602	5.2	0.688	**1.6**	0.665	4.5

Datasets: We evaluate CIFAR-100 (Krizhevsky, 2009), ImageNet (Deng et al., 2009), ImageNet-V2 (Recht et al., 2019), and WOS-46985 (Kowsari et al., 2017), and add CIFAR-100-LT, ImageNet-LT and ImageNet-LT-V2: seven settings and 28 dataset-score cells. Following the long-tailed experimental setup of RC3P, LT denotes long-tailed training, the calibration pool and test set are class-balanced. Appendix C.1 details tuning and data splits.
Architectures: Fine-tuned ResNet-50 (He et al. (2016)) for CIFAR. Fine-tuned DistilBERT (Sanh et al. (2019)) on text data. For ImageNet we followed the training procedure of Ding et al. (2023), using SimCLR representations (Chen et al. (2020)) and a single layer prediction head.
Learned Representations: We follow Yosinski et al. (2014) to extract CNN activations and Sanh et al. (2019) for Transformer hidden states.
Baseline methods: We compared our CFCP method to Split, ICP, CCP, NCP and RC3P by evaluating each on these score methods: LAC, APS, RAPS and SAPS.

Evaluation metrics **Class Coverage (COV)** is the fraction of classes whose empirical class-conditional coverage is at least $1-\alpha$. **Efficiency (Set Size)** is the mean number of predicted labels. **Weighted under-coverage (WUC)** averages positive coverage deficits below $1-\alpha$ using empirical class frequencies.
We use COV and set size as the primary criteria, with WUC diagnosing residual undercoverage. COV complements UCR/FracUnderCov used by RC3P/CCP.

Coverage-size evaluation. We first compare coverage-matched operating points, then interpolate each selected baseline's calibration level within each split to CFCP's nominal mean set size. The COV/WUC target remains 0.9. Equal size need not imply equal marginal coverage. Range-average comparisons divide the area under each metric-size curve by the shared interval width. LT intervals are specific to each CFCP-baseline pair and supported on all ten splits. We retain the matched-budget comparator for the corresponding range-average comparison. Paired differences use ten splits and Student-t 95% confidence intervals. An interval containing zero is inconclusive. Coverage-matched tables use five splits. Appendix C.3 states the matching protocol in full.

4.1 Results

Coverage-matched comparisons as an initial diagnostic. On the four original settings, CFCP attains the highest mean COV in 8 of 16 cells, counting ties at displayed precision, and has smaller or tied mean set size than the highest-COV baseline in 14 (Table 1). These results motivate evaluating classwise coverage at equal size budgets. Figure 3 illustrates the coverage-size tradeoff. Over the 12 matched sizes per score, the reported interval rule favors CFCP at 9/12 points under LAC and

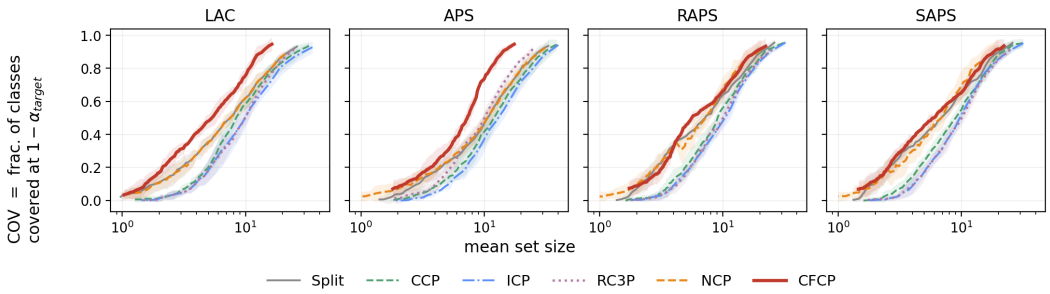


Figure 3: **Empirical coverage-size frontier on CIFAR-100.** COV against mean set size, sweeping the calibration α over $[0.03, 0.40]$ while keeping the classwise evaluation target at 0.9. Bands are 95% CIs over 5 seeds.

6/12 under APS, with no point favoring the selected baseline under that rule. The following paired comparisons quantify it across all seven settings.

Table 2: Paired per-split difference in class coverage, $\Delta\text{COV} = \text{COV}_{\text{CFCP}} - \text{COV}_{B}$, against the strongest baseline in each cell at matched efficiency. For every split the baseline's calibration level is interpolated to the mean set size CFCP produces there, so the two sides spend the same set-size budget. On CIFAR-100, ImageNet, ImageNet-V2 and WOS-46985 CFCP is likewise read off its own alpha-swept curve at that size, and on the long-tailed rows at its own operating point. Both methods are read on the same ten calibration/test splits, so the difference is formed within each split and the ten paired differences are then summarized as mean $\pm$ Student-t 95% CI, in percentage points. Positive values favor CFCP. **Bold** marks a cell whose interval is in CFCP's favor, underline marks one whose interval is in the baseline's favor. The curves are those of Table 3. Appendix Table 9 reports the two methods separately at the same matched point but from the point-alpha sweep, a different set of fits, so the difference of its columns is not this ΔCOV.

Dataset	ΔCOV (pp)			
	LAC	**APS**	**RAPS**	**SAPS**
CIFAR-100	$\mathbf{+14.62 \pm 2.62}$	$\mathbf{+20.86 \pm 3.55}$	$\mathbf{+4.21 \pm 2.14}$	$\mathbf{+2.72 \pm 2.56}$
CIFAR-100-LT	$\mathbf{+14.83 \pm 3.34}$	$\mathbf{+22.47 \pm 4.78}$	$\mathbf{+6.73 \pm 2.22}$	$\mathbf{+6.60 \pm 2.69}$
ImageNet	$\mathbf{+1.37 \pm 0.32}$	$\mathbf{+15.26 \pm 0.64}$	-0.95 ± 0.65	$+0.02 \pm 0.62$
ImageNet-V2	$\mathbf{+1.14 \pm 0.38}$	$\mathbf{+16.01 \pm 0.21}$	-0.33 ± 0.36	$\mathbf{+0.60 \pm 0.42}$
ImageNet-LT	$\mathbf{+1.07 \pm 0.88}$	$\mathbf{+18.20 \pm 0.84}$	$\mathbf{+3.33 \pm 0.68}$	$\mathbf{+1.89 \pm 0.35}$
ImageNet-LT-V2	-0.08 ± 0.44	$\mathbf{+15.25 \pm 0.45}$	$\mathbf{+1.86 \pm 0.33}$	$\mathbf{+0.91 \pm 0.40}$
WOS-46985	$+0.78 \pm 0.94$	$\mathbf{+2.50 \pm 1.06}$	$\mathbf{+2.39 \pm 1.24}$	-1.47 ± 1.59

Matched baseline: CIFAR-100: NCP (all). CIFAR-100-LT: NCP (all). ImageNet: Split (LAC), NCP (APS), NCP (RAPS), NCP (SAPS). ImageNet-V2: NCP (LAC), NCP (APS), NCP (RAPS), Split (SAPS). ImageNet-LT: NCP (all). ImageNet-LT-V2: NCP (all). WOS-46985: NCP (all). Decided cells (CFCP/baseline/undecided): 22/1/5.

Class coverage at matched size. At CFCP's nominal size budget, paired COV comparisons favor CFCP in 22 of 28 cells and the selected baseline in one, with five inconclusive (Table 2). CFCP is favored in 11 of the 12 LT cells, ImageNet-LT-V2/LAC is inconclusive. APS favors CFCP in all seven settings, with LT gains of 22.47 ± 4.78, 18.20 ± 0.84, and 15.25 ± 0.45 percentage points on CIFAR-100-LT, ImageNet-LT, and ImageNet-LT-V2, respectively. The only decided baseline advantage is ImageNet/RAPS at 0.95 ± 0.65 points. NCP is the strongest baseline in 26 of 28 cells.

Matching size changes marginal coverage. Moving the selected baseline to CFCP's budget lowers its marginal coverage in 25 of 28 cells, including all 12 LT cells. 22 of these decreases have intervals excluding zero (Appendix Table 7). The largest decrease is 15.94 ± 0.84 points on CIFAR-100-LT/APS. Thus, the COV gains measure performance at equal mean size, where marginal coverage can differ.

Table 3: Paired per-split range-average difference in class coverage, i.e. the paired difference in area under the COV vs. mean-set-size curve divided by the width of the cell's set-size window, in percentage points. Each cell integrates both methods over the same set-size window, the range every method's own curve covers on every split, so the comparison holds across the whole coverage-efficiency trade-off rather than at one matched operating point. Both methods are read on the same ten calibration/test splits, so the difference is formed within each split and the ten paired differences are summarized as mean ± Student-t 95% CI. Positive values favor CFCP. **Bold** marks a cell whose intervals are in CFCP's favor, underline marks one whose in the baseline's favor

Dataset	$\Delta\mathrm{AUC}_{\mathrm{COV}}$ (pp)			
	LAC	**APS**	**RAPS**	**SAPS**
CIFAR-100	$\mathbf{+13.48 \pm 1.40}$	$\mathbf{+18.39 \pm 1.39}$	$\mathbf{+1.46 \pm 1.03}$	$\mathbf{+1.44 \pm 1.41}$
CIFAR-100-LT	$\mathbf{+18.11 \pm 0.93}$	$\mathbf{+21.67 \pm 1.54}$	$\mathbf{+10.67 \pm 1.18}$	$\mathbf{+10.40 \pm 1.13}$
ImageNet	$\mathbf{+0.98 \pm 0.08}$	$\mathbf{+12.85 \pm 0.39}$	$\underline{-0.78 \pm 0.25}$	$\underline{-0.72 \pm 0.33}$
ImageNet-V2	$\mathbf{+1.98 \pm 0.08}$	$\mathbf{+12.18 \pm 0.07}$	$\underline{-0.83 \pm 0.24}$	$\mathbf{+1.15 \pm 0.08}$
ImageNet-LT	$\mathbf{+4.98 \pm 0.26}$	$\mathbf{+14.82 \pm 0.37}$	$\mathbf{+2.35 \pm 0.30}$	$\mathbf{+2.09 \pm 0.33}$
ImageNet-LT-V2	$\mathbf{+3.62 \pm 0.15}$	$\mathbf{+13.02 \pm 0.10}$	$\mathbf{+0.88 \pm 0.16}$	$\mathbf{+0.66 \pm 0.16}$
WOS-46985	$\mathbf{+1.46 \pm 0.61}$	$\mathbf{+3.95 \pm 0.88}$	$\mathbf{+1.45 \pm 1.02}$	-0.40 ± 0.65

Matched baseline: CIFAR-100: NCP (all). CIFAR-100-LT: NCP (all). ImageNet: Split (LAC), NCP (APS), NCP (RAPS), NCP (SAPS). ImageNet-V2: NCP (LAC), NCP (APS), NCP (RAPS), Split (SAPS). ImageNet-LT: NCP (all). ImageNet-LT-V2: NCP (all). WOS-46985: NCP (all). Decided cells (CFCP/baseline/undecided): 24/3/1.

Performance over shared size ranges. Range-average COV favors CFCP in 24 of 28 cells and the selected baseline in three, with one inconclusive (Table 3). All 12 LT comparisons favor CFCP, with gains from 0.66 ± 0.16 to 21.67 ± 1.54 points. The three decided baseline advantages occur on ImageNet/RAPS, ImageNet/SAPS, and ImageNet-V2/RAPS. ImageNet-V2/SAPS favors CFCP over the shared range by 1.15 ± 0.08 points. Average gains therefore support a broader tradeoff advantage without implying dominance at every budget.

Residual undercoverage. Paired WUC comparisons favor CFCP in 20 of 28 cells at matched size and 20 of 28 over shared size ranges, the selected baseline is favored in three and five, respectively (Appendix C.4). All 12 LT comparisons favor CFCP under both evaluations. The remaining five matched-size and three range-average comparisons are inconclusive. These results complement COV, but aggregate improvements can coexist with larger deficits for individual classes.

CFCP-NCP Comparison. The direct comparison with NCP, the strongest matched-size baseline in most cells, evaluates two localization strategies: NCP weights neighboring calibration scores to obtain a query-specific threshold, whereas CFCP adjusts class probabilities using cluster-label frequencies before global split-conformal calibration. The comparison thus concerns the complete strategies, which differ in both the stage of localization and the construction of local information. At matched size across the four original settings, paired COV favors CFCP in 11 of 16 cells and NCP in one, with four inconclusive (Appendix Table 6). The reported prediction times on CIFAR-100 and WOS-46985 are lower for CFCP, with approximately 5-35$\times$ speedups over the evaluated NCP implementation. This comparison concerns the conformal procedure with cached model outputs. The structural saving comes from querying a compact centroid representation. The full CFCP query also includes probability mixing and score construction. Appendix C.2 describes the timing accounting, including the reuse of NCP searches on ImageNet.

Ablation results. CFCP has two separable components: the substitution of the cluster-frequency mixture q_x^{local} for the base model's π_x, and the reliability weight $r(x)$ that interpolates between them. We ablate each by fixing r: $r \equiv 0$ recovers π_x exactly and reduces CFCP to the base-model conformal predictor, while $r \equiv 1$ uses the cluster mixture everywhere, retaining cluster-level prior smoothing but no fallback to π_x. All three arms are scored through the same set constructor, each with its own split-conformal threshold, at the nominal $\alpha = 0.1$ over the reported repeated splits (Table 12, Appendix C.5).

CLUSTER FREQUENCIES. On CIFAR-100 the full method dominates π_x, covering at least as much with a set no larger in the same split in 17 of 20 split $\times$ score cells, and reduces set size by 10.2% (SAPS) to 24.4% (APS). The gain is not confined to the set constructor: as a probability vector, the cluster mixture alone improves the per-sample negative log-likelihood over π_x by 0.58 ± 0.07, nor is it bought with coverage: moving π_x along its own coverage-size curve to CFCP's realized coverage - the full method's sets are 12.0% to 37.9% smaller in every CIFAR-100 split.

RELIABILITY WEIGHT. $r(x)$ is designed to use the cluster evidence where it is informative and fall back to π_x where it is not, and the ablation shows both behaviours. On CIFAR-100 the interpolation is better than either endpoint: its NLL gain over π_x is 0.75 ± 0.06, against 0.58 ± 0.07 for $r \equiv 1$. On WOS-46985, where the cluster mixture is a worse probability estimate than π_x (-0.21 for $r \equiv 1$), $r(x)$ recovers 84% of that loss (-0.03), limiting the damage that pooling over uninformative clusters would otherwise cause. In the conformal sets, at coverage matched to CFCP's, the weight reduces set size relative to $r \equiv 1$ by 14.3% under APS on CIFAR-100 and by 3.8% under LAC on WOS-46985, the dataset on which the cluster mixture is least reliable in every split of both cells. Three further cells are statistically tied, and under RAPS the unshrunk mixture is smaller ($+6.8\%$ for CFCP). Appendix C.5 gives the matched-coverage protocol, per-split dominance counts, confidence intervals and cluster-level coverage dispersion.

5 DISCUSSION, LIMITATIONS AND FUTURE WORK

CFCP uses local label composition in learned representations to construct probability estimates for standard conformal prediction. Across the evaluated image and text settings, the results support the central hypothesis that prediction-set efficiency can provide headroom for improved classwise coverage within a size budget. Comparisons at matched mean set size and over shared size ranges favor CFCP in most dataset-score settings, including those with long-tailed training. These findings support adapting the probabilities supplied to conformal scores as a practical approach to improving the classwise coverage-size tradeoff. Independent final calibration preserves marginal validity under exchangeability, while the observed classwise benefits remain empirical.

Limitations. CFCP does not guarantee local or classwise coverage, and the aggregate gains do not imply improvements for every class or at every size budget. Our main classwise metrics condition on the true label, whereas coverage within an input-defined representation region is a different target. A higher fraction of classes above the empirical coverage target can coexist with larger deficits in a smaller set of classes, as illustrated in Appendix C.6.

The method relies on useful representation-space structure. Pooling across heterogeneous regions and shrinkage toward an inaccurate prior can introduce bias. The reliability rule is a heuristic, and the reported ablations do not establish uniformly better probability estimation or per-cluster coverage. Both ImageNet-V2 evaluations reuse their respective ImageNet fits and calibration under distribution shift, so the exchangeability-based guarantee does not automatically apply.

The LT experiments follow the experimental setup of RC3P to study imbalanced training with balanced calibration and test populations, rather than long-tailed test prevalence. They also estimate local label frequencies from a balanced calibration pool using Protocol B (Appendix C.1.2). These results therefore evaluate the complete fitting protocol and do not isolate the effect of clustering from the change in label information available for probability estimation.

Future work. Useful extensions include class or cluster-aware calibration to address residual undercoverage, further evaluation of the reliability rule, and testing the local-frequency construction with other representations and data modalities. Adapting the framework to regression would require a suitable local predictive summary and interval construction.

6 STATEMENTS

6.1 AI USE STATEMENT

This research used generative AI tools assistance in programming the method implementation and the empirical evaluation, suggest experimental parameters. In addition, Generative AI tools assisted with preparation of this manuscript, including prose drafting, LaTeX construction and checks against the available implementation and recorded results. We have not used generative AI tools for the brainstorming or ideation or development of the method and its core concepts. We have not used AI tools for a background research or related work. We have reviewed all AI-assisted work: AI-generated prose and code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

6.2 ETHICS STATEMENT

The research conforms with the ICLR Code of Ethics. The paper uses only publicly available benchmark datasets, does not involve human subjects, and the proposed method is a general-purpose improvement to conformal prediction that does not raise specific ethical concerns.

6.3 REPRODUCIBILITY STATEMENT

The method is specified in Section 2, with the score construction, the reliability rule and the LAC/APS/RAPS/SAPS instantiations in Appendix A. All assumptions and complete proofs of the theoretical claims are given in Appendix B. Appendix C.1 describes the datasets, their splits, the base models and their training recipes, the representations used, the metric definitions, and the hyperparameters of CFCP and of every baseline. The coverage-matched, matched-size and coverage-size-range protocols are described in Section 4 and Appendices C.2, C.3 and C.4, and the ablation protocol in Appendix C.5. Anonymous code is available at `https://anonymous.4open.science/r/CFCP-C039/`. It contains our implementation of CFCP, our implementations of the NCP and RC3P baselines, a single driver that evaluates CFCP and all baselines over repeated calibration/test splits, and the hyperparameter configuration files used for the reported cells. Because the conformal procedures operate on cached model outputs, the reported numbers can be regenerated without retraining. The accompanying README documents, for each dataset, the cached arrays required, the row order they must follow, and the per-seed split rule, so the released splits match the ones used here, together with the fixed seeds and the remaining sources of run-to-run variation.

REFERENCES

Anastasios Angelopoulos, Stephen Bates, Jitendra Malik, and Michael I Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in Neural Information Processing Systems*, 33:22243–22255, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Tiffany Ding, Anastasios N. Angelopoulos, Stephen Bates, Michael I. Jordan, and Ryan J. Tibshirani. Class-conditional conformal prediction with many classes. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2306.09335. NeurIPS 2023.

Tiffany Ding, Jean-Baptiste Fermanian, and Joseph Salmon. Conformal prediction for long-tailed classification. *arXiv preprint arXiv:2507.06867*, 2025.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *IEEE Transactions on Big Data*, 2025.

Subhankar Ghosh, Taha Belkhouja, Yan Yan, and Janardhan Rao Doppa. Improving uncertainty quantification of deep classifiers via neighborhood conformal prediction: Novel algorithm and theoretical analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 7722–7730, 2023. doi: 10.1609/aaai.v37i6.25936. URL https://ojs.aaai.org/index.php/AAAI/article/view/25936.

Natalia Martinez Gil, Dhaval Patel, Chandra Reddy, Giri Ganapavarapu, Roman Vaculin, and Jayant Kalagnanam. Identifying homogeneous and interpretable groups for conformal prediction. In *Uncertainty in Artificial Intelligence*, pp. 2471–2485. PMLR, 2024.

Leying Guan. Localized conformal prediction: A generalized inference framework for conformal prediction. *Biometrika*, 110(1):33–50, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Rohan Hore and Rina Foygel Barber. Conformal prediction with local weights: Randomization enables local guarantees. *arXiv preprint arXiv:2310.07850*, 2023. URL https://arxiv.org/abs/2310.07850.

Jianguo Huang, Huajun Xi, Linjun Zhang, Huaxiu Yao, Yue Qiu, and Hongxin Wei. Conformal prediction for deep classifier via label ranking. *arXiv preprint arXiv:2310.06430*, 2023.

Jianguo Huang, Jianqing Song, Xuanning Zhou, Bingyi Jing, and Hongxin Wei. Torchcp: A python library for conformal prediction, 2024.

Kamran Kowsari, Donald E Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S Gerber, and Laura E Barnes. Hdltex: Hierarchical deep learning for text classification. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pp. 364–371. IEEE, 2017.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Yuanjie Shi, Subhankar Ghosh, Taha Belkhouja, Janardhan R Doppa, and Yan Yan. Conformal prediction for class-wise coverage via augmented label rank calibration. *Advances in Neural Information Processing Systems*, 37, 2024.

Vladimir Vovk. Conditional validity of inductive conformal predictors. In Steven C. H. Hoi and Wray Buntine (eds.), *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pp. 475–490, Singapore, 2012. PMLR. URL https://proceedings.mlr.press/v25/vovk12.html.

Vladimir Vovk and Ivan Petej. Venn-abers predictors. *arXiv preprint arXiv:1211.0025*, 2012.

Vladimir Vovk, Glenn Shafer, and Ilia Nouretdinov. Self-calibrating probability forecasting. *Advances in neural information processing systems*, 16, 2003.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, 27, 2014.

Yao Zhang and Emmanuel J. Candès. Posterior conformal prediction. *arXiv preprint arXiv:2409.19712*, 2024. URL https://arxiv.org/abs/2409.19712.

Table 4: Notation used in the CFCP method.

Symbol	Type	Definition
$x \in \mathcal{X}$	input	Input example.
$y, c \in \mathcal{Y}$	label	True label y or candidate class c.
C	scalar	Number of classes, $\lvert\mathcal{Y}\rvert = C$.
f	model	Pre-trained classifier.
$\phi(x) \in \mathbb{R}^d$	vector	Learned representation of x.
z	vector	ℓ_2-normalized embedding of x.
d	scalar	Embedding dimension.
K	scalar	Number of representation-space clusters.
$\mu_k \in \mathbb{R}^d$	vector	Centroid of cluster k, for $k = 1, \ldots, K$.
m	scalar	Number of nearest centroids used for soft assignment.
$\kappa_j(x)$	index	Index of the j-th nearest centroid to x.
$\delta_j(x)$	scalar	Cosine similarity between z and centroid $\mu_{\kappa_j(x)}$.
$w_j(x)$	scalar	Soft assignment weight of x to its j-th nearest centroid.
τ	scalar	Softmax temperature controlling assignment sharpness. Smaller values yield more local assignments.
$\mathcal{D}_{\text{freq}}$	set	Subset used to estimate cluster-level label statistics.
$\mathcal{D}_q$	set	Held-out subset used to estimate the conformal threshold.
N_{kc}	scalar	Soft count of frequency-sample observations from class c assigned to cluster k.
S_k	scalar	Effective support of cluster k, i.e., total soft mass assigned to cluster k.
$\pi \in \Delta^{C-1}$	vector	Global prior used for smoothing cluster label distributions.
β	scalar	Dirichlet smoothing strength toward π.
$\hat{p}_k(c)$	scalar	Smoothed empirical probability of class c in cluster k.
$q_x^{\text{local}}(c)$	scalar	Local cluster-frequency mixture probability for class c at test point x.
$\pi_x(c)$	scalar	Fallback prior for class c at x, e.g., the base model probability.
$\eta(x)$	vector	The conditional label law.
β_{sup}	scalar	Support-shrinkage parameter controlling the effect of neighborhood support.
$r(x) \in [0, 1]$	scalar	Reliability weight used to interpolate between local cluster frequencies and the fallback prior.
$q_x(c)$	scalar	Final CFCP probability assigned to class c for input x.
$s(x, c)$	scalar	Nonconformity score of candidate class c for input x.
α	scalar	Target miscoverage level.
$\widehat{t}_\alpha$	scalar	Split-conformal quantile calibrated on $\mathcal{D}_q$.
$\hat{\Gamma}(x)$	set	Conformal prediction set returned for input x.

A Additional Method Details

A.1 CFCP score construction details

Let the embeddings and centroids be ℓ_2-normalized. We use the neighbor indices $\kappa_j(x)$, weights $w_j(x)$, and cluster-indexed assignments $a_k(x)$ from equation 1. In particular, $a_{ik} = \sum_{j=1}^{m} w_j(x_i)\mathbf{1}\{\kappa(x_i) = k\}$. The counts $N_{kc} = \sum_i a_{ik}\mathbf{1}\{y_i = c\}$ and supports $S_k = \sum_i a_{ik}$ are accumulated only from D_{freq}, as in equation 2. A zero-support cluster uses the prior because $\beta > 0$.

For $m > 1$, the default temperature is

$$\tau = \frac{\text{median}_{x \in D_{\text{ref}}}\{\delta_1(x) - \delta_m(x)\}}{\log m},$$

with a small positive numerical floor when the median gap is zero. Here D_{ref} is the reference embedding set used for the geometric default, taken from the training pool. For $m = 1$, the single assignment weight is one and the temperature has no effect. An explicit temperature may also be supplied.

The smoothing prior π is either the empirical label prior or the mean of the base model's class-probability vectors on D_{freq}. The default smoothing strength is

$$\beta = \frac{\rho}{1-\rho} S_{\mathrm{typ}}, \qquad S_{\mathrm{typ}} = \mathrm{median}_{k:S_k>0}\, S_k, \qquad \rho = 0.1.$$

Thus a cluster with typical support assigns fraction ρ of its smoothed distribution to the prior. The cluster distributions and query mixture are those in equation 3.

A.2 Reliability shrinkage details

The final vector is $q_x = r(x) q_x^{\mathrm{local}} + (1 - r(x))\pi_x$. In the main formulation, π_x is the base model's sample-specific prediction. The implementation also supports a constant empirical label prior or mean model prediction as the fallback. These choices should be distinguished when reporting configurations: only the base-model fallback gives $q_x = \pi_x$ as the original model prediction when $r = 0$.

The reliability weight combines assignment concentration and weighted cluster support:

$$\mathrm{conf}(x) = (\max_j w_j(x))^{\gamma}, \qquad \sup(x) = \sum_j w_j(x) s_j, \qquad r(x) = \mathrm{conf}(x) \frac{\sup(x)}{\sup(x) + \beta_{\mathrm{sup}}}.$$

The default is $\gamma = 1$. The support parameter is

$$\beta_{\mathrm{sup}} = \frac{\rho_{\mathrm{sup}}}{1-\rho_{\mathrm{sup}}} \mathrm{median}_{x \in D_{\mathrm{freq}}} \sup(x), \qquad \rho_{\mathrm{sup}} = 0.5,$$

so the default equals the median weighted support. Parameters may also be overridden using development data.

These choices reduce the local mixture's weight for diffuse assignments or low support. Because the smoothing defaults scale with support, their contributions need not vanish as more frequency data are collected. Accordingly, the oracle substitution $q_x = \eta(x)$ in Theorem 2 is not asserted to be an asymptotic limit of this implementation.

A.3 LAC/APS/RAPS/SAPS instantiations

The final calibration sample $D_q = \{(x_i, y_i)\}_{i=1}^{n_q}$ is used for the conformal threshold. For a selected score s, including its auxiliary randomization if present, compute $s_i = s(x_i, y_i, U_i)$ and use the order statistic in equation 5, with the $+\infty$ convention when the required rank is $n_q + 1$.

Use a fixed deterministic tie rule to sort labels as $\sigma_x(1), \ldots, \sigma_x(C)$ in decreasing order of q_x, and write $Q_j(x) = q_x(\sigma_x(j))$. For a candidate label $y = \sigma_x(j)$:

- LAC uses $s_{\mathrm{LAC}}(x, y) = 1 - q_x(y)$.
- Randomized APS uses $s_{\mathrm{APS}}(x, y, u) = \sum_{\ell<j} Q_\ell(x) + u Q_j(x)$, where $u \sim \mathrm{Unif}(0, 1)$. The deterministic version uses $u = 1$.
- RAPS adds $\lambda(j - k_{\mathrm{reg}})_+$ to the APS score, with $\lambda \geq 0$.
- Randomized SAPS uses $Q_1(x)(1-u)$ for $j = 1$ and $Q_1(x) + \omega(j - 1 - u)$ for $j \geq 2$, with $\omega > 0$. Its deterministic version uses the right endpoints of these intervals.

The raw set contains all labels whose score does not exceed the calibrated threshold. For these score conventions and a scalar threshold it is the longest admissible prefix of the sorted labels, possibly empty. The implementation adds the top-ranked label when needed to ensure nonemptiness. Calibration and prediction use the same true-label score distribution. Appendix B.1 details the randomization conventions.

A.4 Computational complexity

CFCP uses FAISS (Douze et al., 2025) for spherical k-means and retrieval. A full-batch clustering run has cost $O(Tn_{\text{train}}Kd)$. After clustering, accumulating the soft counts requires centroid similarities for each frequency example, updates for its m retrieved clusters and initialization/smoothing of the $K \times C$ count table. These components cost $O(n_{\text{freq}}Kd + n_{\text{freq}}m + KC)$, apart from neighbor selection.

At prediction time, the similarity calculation over K centroids costs $O(Kd)$, and mixing m dense class distributions costs $O(mC)$. Top-m selection adds an implementation-dependent cost: linear selection is possible, while a heap-based implementation can require $O(K \log m)$ for $m > 1$. LAC set construction costs $O(C)$. Sorting for the reported APS/RAPS/SAPS implementations costs $O(C \log C)$. Storage for the fitted local model is $O(Kd + KC)$, in addition to the base predictor and batch workspace.

For fixed K, m, C, d, this fitted-model prediction cost is independent of the number of frequency examples. By comparison, an exhaustive NCP query evaluates similarities to n_{cal} calibration examples, at cost $O(n_{\text{cal}}d)$, and computes a localized quantile. Approximate search, batching, and reuse affect practical timings, so this comparison is not a lower bound on every possible neighborhood implementation. The statistical tradeoff is evaluated separately in the experiments.

B Theoretical analysis: assumptions and proofs

B.1 Statistical setting and score conventions

Let $(X, Y) \sim P$, with $Y \in \{1, \dots, C\}$ and $\eta_c(x) = \Pr\{Y = c \mid X = x\}$. Prediction randomization is independent of the unknown test label conditional on the input. The fitted model $\widehat{M}$ contains the representation, centroids, frequencies, priors, reliability rule, score family, and all hyperparameters.

Fitting and final calibration. A sufficient protocol for Theorem 1 fits and tunes these objects on data independent of D_q and the test point, followed by i.i.d. calibration and test examples from P. The theorem itself only requires exchangeability conditional on $\widehat{M}$. Merely avoiding calibration labels does not establish this property for an arbitrary procedure that uses calibration inputs.

The frequency sample may equal the model-training sample: the marginal proof conditions on the entire fitted score and makes no frequency-estimation accuracy or concentration assumption. The second result instead studies the idealized population substitution $q_x = \eta(x)$.

Calibration threshold. Fix $\alpha \in (0, 1)$ before final calibration, and suppose $n = n_q \geq 1$. For calibration scores $s_1, \dots, s_n$, define

$$\widehat{t}_\alpha = \begin{cases} s_{(k_\alpha)}, & k_\alpha \leq n, \\ +\infty, & k_\alpha = n+1, \end{cases} \qquad k_\alpha = \lceil (n+1)(1-\alpha) \rceil, \quad \nu_{n,\alpha} = \frac{k_\alpha}{n+1}. \tag{5}$$

Here $s_{(j)}$ is the j-th order statistic of the calibration scores. In particular, the rank is not clamped to n. We have $1 - \alpha \leq \nu_{n,\alpha} \leq 1$ and $\nu_{n,\alpha} - (1-\alpha) < 1/(n+1)$. Any data-dependent choice of the score or calibration level must be completed before this final calibration sample is used, unless an additional validity argument is supplied.

Score definitions. Fix a measurable deterministic tie rule for sorting q_x, write $\sigma_x(1), \dots, \sigma_x(C)$ for the sorted labels, and let $Q_j(x) = q_x(\sigma_x(j))$. If $y = \sigma_x(j)$, the scores are

$$s_{\text{LAC}}(x, y) = 1 - q_x(y), \tag{6}$$

$$s_{\text{APS}}(x, y, u) = \sum_{\ell < j} Q_\ell(x) + u Q_j(x), \qquad u \sim \text{Unif}(0, 1), \tag{7}$$

$$s_{\text{RAPS}}(x, y, u) = s_{\text{APS}}(x, y, u) + \lambda(j - k_{\text{reg}})_+, \qquad \lambda \geq 0, \tag{8}$$

$$s_{\text{SAPS}}(x, y, u) = \begin{cases} Q_1(x)(1-u), & j = 1, \\ Q_1(x) + \omega(j - 1 - u), & j \geq 2, \end{cases} \quad \omega > 0. \tag{9}$$

The SAPS convention matches the score distribution of the CFCP implementation. Calibration and prediction must use the same true-label score law. One uniform may be shared across labels for APS/RAPS. Independent uniforms across labels are also permitted for the marginal argument, and are used by the implementation's SAPS prediction routine. Randomization is independent across calibration and test examples. The deterministic APS/RAPS versions use $u = 1$, while deterministic SAPS uses the right endpoints of its score intervals. Their marginal lower bound remains valid.

Raw and nonempty prediction sets. With $\mathbf{U}$ denoting the appropriate prediction randomization, set

$$\widehat{\Gamma}(x;\mathbf{U}) = \{c : s(x,c;\mathbf{U}) \le \widehat{t}_\alpha\}, \qquad \widehat{\Gamma}^+(x;\mathbf{U}) = \begin{cases} \widehat{\Gamma}(x;\mathbf{U}), & \widehat{\Gamma}(x;\mathbf{U}) \neq \varnothing, \\ \{\sigma_x(1)\}, & \widehat{\Gamma}(x;\mathbf{U}) = \varnothing. \end{cases} \tag{10}$$

This is the nonempty-set convention of the marginal CFCP implementation. With a scalar threshold, each supported score is nondecreasing in sorted rank, so the raw set is a prefix. Thus the nonempty rule is also equivalent to taking its union with a top-ranked label. This section analyzes scalar thresholds, it does not assume that a class-specific threshold vector produces a prefix.

B.2 Marginal validity and the effect of nonemptiness

Proof of Theorem 1. Condition on $\widehat{M}$. The score map is fixed, and appending independent randomization variables with the same distribution to exchangeable calibration and test examples preserves exchangeability. The $n+1$ true-label scores are consequently exchangeable. For purposes of the rank argument only, break any score ties using independent continuous tie breakers. The test score's resulting rank is uniform on $\{1,\dots,n+1\}$. If this rank is at most $k_\alpha \le n$, then its unperturbed score is at most the k_α-th calibration score. Hence

$$\Pr\{s_{n+1} \le \widehat{t}_\alpha \mid \widehat{M}\} \ge \frac{k_\alpha}{n+1} = \nu_{n,\alpha}.$$

When $k_\alpha = n+1$, the threshold is $+\infty$ and the set contains all labels. Finally, $\widehat{\Gamma}(x) \subseteq \widehat{\Gamma}^+(x)$ pointwise, so set enlargement preserves the lower bound. No continuity assumption is needed. □

Effect of the nonempty rule. Let $E = \{\widehat{\Gamma}(X) = \varnothing\}$. Pointwise, the nonempty rule adds one label exactly on E, and adds a correct label exactly on $E \cap \{Y = \sigma_X(1)\}$. Thus, conditional on $\widehat{M}$,

$$\Pr\{Y \in \widehat{\Gamma}^+(X) \mid \widehat{M}\} = \Pr\{Y \in \widehat{\Gamma}(X) \mid \widehat{M}\} + \Pr\{E, Y = \sigma_X(1) \mid \widehat{M}\}, \tag{11}$$

$$\mathbb{E}[|\widehat{\Gamma}^+(X)| \mid \widehat{M}] = \mathbb{E}[|\widehat{\Gamma}(X)| \mid \widehat{M}] + \Pr\{E \mid \widehat{M}\}. \tag{12}$$

Set enlargement preserves the lower coverage guarantee but changes expected size and can introduce additional conservativeness. We therefore distinguish the implemented nonempty rule from the unrestricted population oracle in Theorem 2.

B.3 Classical population oracle optimality for LAC

Proof of Theorem 2. We give the standard threshold argument underlying the least-ambiguous classifier characterization of Sadinle et al. (2019). Represent an arbitrary randomized predictor by its conditional inclusion probabilities

$$h_c(x) = \Pr\{c \in \Gamma(x) \mid X = x\} \in [0,1].$$

Independence of the inclusion decisions across labels is not required. Linearity gives

$$\mathbb{E}|\Gamma(X)| = \mathbb{E}\sum_c h_c(X), \qquad \Pr\{Y \in \Gamma(X)\} = \mathbb{E}\sum_c \eta_c(X) h_c(X).$$

There exist $\tau_\alpha > 0$ and $\xi \in [0,1]$ such that

$$h_c^*(x) = \mathbf{1}\{\eta_c(x) > \tau_\alpha\} + \xi\,\mathbf{1}\{\eta_c(x) = \tau_\alpha\} \tag{13}$$

has coverage $1-\alpha$. Indeed, for $W = \eta_Y(X)$ we have $W > 0$ almost surely. Choose a positive threshold with $\Pr\{W > \tau_\alpha\} \le 1-\alpha \le \Pr\{W \ge \tau_\alpha\}$, randomization on its atom supplies any missing coverage.

For any competitor h with coverage at least $1-\alpha$, pointwise

$$\{h_c^*(x) - h_c(x)\}\{\eta_c(x) - \tau_\alpha\} \geq 0.$$

Summing and integrating yields

$$\begin{aligned} 0 &\leq \mathbb{E}\sum_c \eta_c(X)\{h_c^*(X) - h_c(X)\} - \tau_\alpha \mathbb{E}\sum_c \{h_c^*(X) - h_c(X)\} \\ &\leq -\tau_\alpha\{\mathbb{E}|\Gamma^*(X)| - \mathbb{E}|\Gamma(X)|\}. \end{aligned}$$

Since $\tau_\alpha > 0$, the claimed minimum-size property follows. When $q_x = \eta(x)$, this rule is precisely a population LAC threshold with boundary randomization. □

C Experiments and results

C.1 Experimental setup

C.1.1 Data sets and models

- **CIFAR-100** (Krizhevsky (2009)): we followed the training procedure from Ding et al. (2023).
  - Splits: The datasets include 50,000 train images and 10,000 test images equally distributed over 100 classes. We split the training data to 60% and 40% validation. The test data was split to repeated holdouts with fixed seeds to 75% calibration, and 25% testing.
  - Model: PyTorch (Paszke et al. (2019)) ResNet-50 (He et al. (2016)), initialized to the IMAGENET1K_V2 (Deng et al. (2009)) weights and then fine-tuned all parameters for 30 epochs. Validation accuracy is 60%.
  - CIFAR-100-LT subsamples the training set to the exponential profile of Cao et al. (2019) at imbalance factor 100 (10,847 images, 500 for the head class down to 5 for the tail class) and repeats the CIFAR-100 fine-tuning recipe on it. Calibration and test are unchanged: the same balanced 7,500/2,500 partition of the CIFAR-100 test set. Accuracy on that balanced pool is 27%.
- **ImageNet** (Deng et al. (2009)): we followed the training procedure from Ding et al. (2023).
  - Splits: The classifier was trained on a random 10% of the ≈1.28M training images, and CFCP's k-means centroids are fitted on a disjoint random subset of 200,000 training images. The 50,000-image ImageNet validation set was split to repeated stratified holdouts with fixed seeds to 50% conformal prediction calibration and 50% testing, 25 images per class each, the same split as ImageNet-LT. In addition, we used ImageNet-V2's test set (10,000 images) with the classification and conformal prediction models of ImageNet's experiment.
  - Why calibration comes from ImageNet val: The SimCLR-v2 backbone was trained on the whole ImageNet training set, so it has seen every training image and no validation image, and the two sets were collected separately. Calibrating on training images and testing on validation images is therefore not exchangeable: held-out training images are easier for the model (top-1 accuracy 77.6% against 73.8% on the validation set), and split conformal then under-covers by an amount that depends on the score. With calibration and test both drawn from the validation set, split conformal attains the nominal coverage for every score.
  - Model: Based on a SimCLR-v2 model (Chen et al. (2020)), which was trained on the ImageNet training set without labels, to extract feature vectors of length 6,144 for all images in the ImageNet training set, we trained a classifier using one fully connected prediction head. Top-1 accuracy on the ImageNet validation set is 73.8%.
  - ImageNet-LT follows the official per-class training counts using sampled feature rows. Calibration and test are a per-seed stratified 50/50 split of the ImageNet validation set, 25 images per class each, rather than a sample of the non-LT training rows, so both are class-balanced. ImageNet-LT-V2 evaluates the same LT-trained model and the same calibration on ImageNet-V2.

- **WOS-46985**: We used the keywords and abstract texts to train a classification model on the dataset.
  - Splits: 70% of the data were randomly sampled for training. The remaining 30% were split to repeated holdouts with fixed seeds to 65% conformal prediction calibration and 35% testing.
  - Model: We used DistilBERT (Sanh et al. (2019)) classifier fine-tuned with num_train_epochs=5, learning_rate=2e-5, weight_decay=0.01, warmup_ratio=0.1. Validation accuracy is 88.4%.
- **Baseline methods**: We compared our CFCP method to Split, ICP, CCP, NCP and RC3P by evaluating each on these score methods: LAC, APS, RAPS and SAPS. For the Split, ICP and CCP, baseline methods and all scores implementation we used the TorchCP framework (Huang et al. (2024)). For the NCP and RC3P baselines we repeated the implementation that was shared by the authors in the original papers.
- **Method parameters**:
  - Score: We used equal parameters for scoring on both CFCP and baseline methods. APS/RAPS/SAPS: randomized=True, RAPS: penalty=0.1, kreg=20 for WOS-46985 and kreg=5 for all other datasets, SAPS: weight=0.2.
  - Prediction set construction: We used temperature=1 for all methods.
  - CCP: ratio_clustering="auto", num_clusters="auto".
  - NCP: we follow the algorithm in the authors' released implementation, which weights the calibration scores by $\exp(-d/\lambda_L)$ and takes an ordinary quantile over the k nearest neighbours. The localizer $\Phi(x)$ is the penultimate representation (CNN activations, Transformer hidden states. For the linear ImageNet head this is the SimCLR feature itself). use_platt_scaling=False, hyper_fraction=0.3, coverage_tolerance=0.01. On a held-out 30% hyper split we select the (λ_L, k) pair with the smallest mean set size among those reaching $1-\alpha-0.01$ coverage, with k over $\{0.05, 0.1, 0.2, 0.4, 0.7, 1.0\}$ of the calibration split and λ_L over $\{0.03, 0.1, 0.3, 1, 3, 10, 30, 100, 10^3, 10^4, 10^5, 10^6\}$ times the observed median k-NN distance. The grids are widened relative to the values stated in the original paper, and λ_L is expressed relative to the data's distance scale, because an absolute grid does not transfer across representations and does not reach the neutral, non-localized end.
  - RC3P: variant = 'mix', mix_grid_size=50, hyper_fraction=0, coverage_slack=0.
  - CFCP hyperparameters were tuned per dataset: n_clusters (K), n_neighbors (m), and, where listed, the assignment temperature τ, the Dirichlet smoothing β, the support shrinkage β_{sup} and the reliability exponent γ. "auto" denotes the geometry- and support-based default of Appendix A. The candidate grids were:
    - CIFAR-100: $K \in \{105, 220, 500, 1000\}$ and $m \in \{40, 80, 160\}$, with $\tau = 0.09375$, $\beta = 0.25$, $\beta_{\text{sup}} = 50$ and $\gamma = 0.25$.
    - CIFAR-100-LT: $K \in \{220\}$, $m \in \{18, 20, 24\}$, $\tau \in \{0.075, 0.08, 0.085\}$, $\beta \in \{0.5, 1\}$, $\beta_{\text{sup}} \in \{10, 25, 40\}$ and $\gamma \in \{0.7, 0.85\}$ (108 configurations).
    - ImageNet: an initial grid $K \in \{300, 1000, 2000\}$, $m \in \{4, 8, 16\}$, $\tau \in \{0.015, 0.03, 0.06, \text{auto}\}$ and $\gamma \in \{0.7, 1\}$ (72 configurations), followed by an extended grid $K \in \{2000, 3000, 4000, 5000\}$, $m \in \{16, 32, 64\}$, $\tau \in \{0.06, \text{auto}\}$ and $\gamma \in \{0.7, 1\}$. β and β_{sup} were fixed at 2 and 25.
    - ImageNet-LT: $K \in \{1000, 2000\}$, $m \in \{8, 16, 32\}$, $\tau \in \{0.03, \text{auto}\}$ and $\gamma \in \{0.7, 1\}$, with $\beta = 2$ and $\beta_{\text{sup}} = 25$ (24 configurations).
    - WOS-46985: $K \in \{64, 128, 256, 512\}$, $m \in \{2, 4, 8, 16, 32\}$, $\tau \in \{0.06, 0.12, 0.24\}$, $\gamma \in \{0.5, 1, 2.5, 5\}$, and $(\beta, \beta_{\text{sup}})$ either both auto or fixed at $(0.5, 50)$, searched in two stages (243 configurations).

    ImageNet-V2 and ImageNet-LT-V2 used the grids of their corresponding ImageNet training setting.
- **Calibration sub-splits**: the standard and long-tailed settings estimate the local cluster-frequency model from different data and therefore reserve different parts of the conformal pool for final calibration. The two protocols are stated in Appendix C.1.2.
- **Coverage**: target $1-\alpha = 0.9$ throughout. Calibration levels are nominal $\alpha = 0.1$ except where Appendix C.3 states otherwise.

- **Uncertainty reporting**: All reported result values are means with 95% CI over repeated dataset splits with fixed seeds.

C.1.2 FITTING PROTOCOLS AND CALIBRATION SUB-SPLITS

CFCP's local model is fixed before final calibration in every experiment, but the data used to estimate it differ between the standard and the long-tailed settings. Both protocols fit the centroids by spherical k-means on the training embeddings, keep the labeled frequency sample D_{freq} disjoint from the final calibration sample D_q, select hyperparameters without reading D_q and never touch the test set. They differ in where D_{freq} comes from and consequently in how much of the conformal pool remains for D_q. The candidate grids for (K, m) are listed above. The selection criterion is stated in the paragraph after Protocol B.

Protocol A: frequencies from the training pool. Used for CIFAR-100, ImageNet, ImageNet-V2 and WOS-46985. The per-cluster label-frequency distributions, their supports and the global priors used for shrinkage are all estimated from the training labels, so $D_{\text{freq}} \subseteq D_{\text{train}}$ and the mixture q_x is a fixed function of the test point that does not depend on the conformal data. No part of the conformal pool is spent on estimating it: for each outer data split the entire non-test pool is used as D_q for the final conformal quantile. Hyperparameter selection likewise never touches D_q or the test set: we partitioned D_{train} into $D_{\text{freq}}, D_{\text{tune}}, D_{\text{eval}}$ with proportions (0.8/0.1/0.1), estimated the cluster frequencies on D_{freq}, calibrated an inner conformal threshold on D_{tune} and scored each candidate (K, m) on the held-out D_{eval} (centroids are fit on the full training pool throughout, as in the final model). For ImageNet this partition is repeated with five inner random seeds, and candidate metrics are averaged before selection. After hyperparameter selection, we refit the CFCP local model on all of D_{train} and calibrated the final conformal threshold on the untouched D_q.

Protocol B: frequencies from the calibration pool. Used for CIFAR-100-LT, ImageNet-LT and ImageNet-LT-V2. Under long-tailed training the label distribution of D_{train} is imbalanced by construction, while the calibration pool and the test set are class-balanced, so cluster label frequencies estimated from training labels would encode the training prior rather than the evaluation one. We therefore draw D_{freq} from the conformal pool. A fixed random partition of that pool, identical across every fit of a given split, reserves 20% as D_q and gives the remaining 80% to the local model: cluster label frequencies, supports and priors. Centroids are still fit on the training embeddings, which carry no labels into q_x. Hyperparameter selection stays inside the 80%: the local model is fit on 60% of the pool, an inner conformal threshold is calibrated on a further 10%, and each candidate (K, m) is scored on the remaining held-out 10%. D_q is the same reserved 20% throughout and is read only for the final threshold. For the ten-split long-tailed runs the candidate metrics are averaged over splits and one configuration per score is fixed before the final fits.

Selection criterion. Every candidate is scored on the held-out slice D_{eval} of its protocol, at the target $\alpha = 0.1$, using the inner threshold from D_{tune}. On D_{eval} we record marginal coverage $\widehat{\text{Cov}}$, mean set size $\bar{S}$, class coverages $\widehat{\text{Cov}}_c$ and the class-coverage dispersion $\text{Disp} = \sum_c \hat{\pi}_c \left(\widehat{\text{Cov}} - \widehat{\text{Cov}}_c\right)_+$, where $\hat{\pi}_c$ is the class frequency on D_{eval} and classes absent from D_{eval} are skipped. Disp measures under-coverage relative to the candidate's own marginal coverage rather than the target, so a candidate cannot lower it by over covering. When a setting has several outer splits or inner replicates, each metric is averaged over them before ranking, and a single winner is selected for all of them.

Size-dispersion rule. A candidate is feasible if $\widehat{\text{Cov}} \geq 1 - \alpha - 0.01 = 0.89$. Across the feasible candidates, Disp and $\bar{S}$ are min-max normalized to $[0, 1]$, and the winner minimizes $\frac{1}{2}\widetilde{\text{Disp}} + \frac{1}{2}\widetilde{S}$. Ties are broken by the smaller $\bar{S}$. If no candidate is feasible, the one with the highest $\widehat{\text{Cov}}$ is taken. The rule is applied separately for each score. Metrics are averaged over the outer splits before ranking.

Matched-size class-coverage rule. The anchor is the D_{eval} mean set size of a fixed reference configuration (K, m) under LAC. Each candidate's $\widehat{\text{COV}}$ on D_{eval} is interpolated in set size to that anchor along its α-curve, then averaged over the inner-split replicates. The winner maximizes this score. Ties are broken by the lower WUC at the anchor. Candidates are scored under LAC only, and the winner is used for all four scores.

Validity under each protocol. Theorem 1 requires the complete fitted model $\widehat{M}$, representation, frequencies, priors and tuning choices, to be fixed before D_q, and the calibration and test scores to be exchangeable conditional on $\widehat{M}$. Protocol A satisfies this with $\widehat{M}$ independent of the entire conformal pool, which is the textbook split-conformal arrangement. Protocol B satisfies it conditionally on the 80% frequency split: given that split, q_x and the score are fixed functions, and the reserved 20% is exchangeable with the test points, so the same bound holds at the reduced calibration size $n_q = 0.2\,|D_{\text{cal}}|$ - 1,500 points on CIFAR-100-LT and 5,000 on ImageNet-LT.

Comparability. Baseline methods were calibrated on the conformal pool of the same outer split, and all methods were evaluated on the same untouched test set. Under Protocol B the baselines calibrate on the full pool while CFCP calibrates on the reserved 20%, so the long-tailed comparisons give CFCP the smaller final calibration sample. Baselines that need an internal split of their own, CCP for its clustering step and NCP for its 30% hyperparameter split, take it from that full pool.

C.1.3 Evaluation Metrics

Class Coverage: We report class coverage, defined as the fraction of classes whose empirical class-conditional coverage meets the target level $1-\alpha$. Let $\widehat{\text{Cov}}_c$ denote the empirical class-conditional coverage for class c. We define Class Coverage at level α as $\widehat{\text{COV}} = \frac{1}{C}\sum_{c=1}^{C} \mathbf{1}\left\{\widehat{\text{Cov}}_c \geq 1-\alpha\right\}$. Higher values are better, and $\widehat{\text{COV}} = 1$ iff all classes satisfy the target coverage level.

Efficiency (Set Size): We measure efficiency by the average set size. Over a test set $\{(x_i, y_i)\}_{i=1}^{n}$ with prediction sets $\widehat{\Gamma}(x_i) \subseteq \{1, \ldots, C\}$, the empirical average set size is $\widehat{\text{Size}} = \frac{1}{n}\sum_{i=1}^{n} |\hat{\Gamma}(x_i)|$. Lower values for identical coverage level indicate greater efficiency.

Weighted under-coverage (WUC): Following the average class coverage gap (CovGap) definition (Ding et al. (2023)), we used its under-coverage variant (UnderCovGap) to evaluate the class conditional coverage. We preferred it over CovGap which penalizes methods which exhibit higher marginal coverage by considering over-coverage as an error. Let $\widehat{\text{Cov}}_c$ denote the empirical class-conditional coverage for class c, and let w_c be the empirical prevalence of class c in the test set, for $c = 1, \ldots, C$ with $\sum_{c=1}^{C} w_c = 1$. We define the weighted under-coverage at level α as $\widehat{\text{WUC}}_p = \sum_{c=1}^{C} w_c \left[(1-\alpha) - \widehat{\text{Cov}}_c\right]_+^p$, where $[u]_+ = \max(u, 0)$ and $p = 1$ in all our reports. This metric measures only coverage deficits relative to the target $1-\alpha$.

C.1.4 Compute

All experiments were performed on a Macbook Pro M4 Max 36GB RAM.

C.2 CFCP-NCP Comparison

C.2.1 Class coverage

C.2.2 Prediction-time cost

The CIFAR-100 and WOS-46985 timings are means over 5 seeds on the same hardware, using identical cached model outputs for both methods. Model forward passes are excluded. Across LAC, APS, RAPS, and SAPS, CFCP takes 0.037-0.047 s on CIFAR-100 versus 0.295-0.301 s for NCP (6.4-8.1×). On WOS-46985 the times are 0.015-0.029 s versus 0.512-0.584 s (17.4-35.3×). On ImageNet they are 1.94-4.09 s versus the reported amortized NCP times of at least 22.4-30.5 s ($\geq 5.5\times$).

The ImageNet NCP entries come from coverage-matching sweeps that share one nearest-neighbor search across four score functions and 11-28 calibration levels. The CFCP entries account for a full CFCP run as reported. These factors describe the measured procedure and reuse policy. An independently optimized implementation or different amortization may yield different ratios.

The structural difference is that CFCP retrieves centroids and mixes their stored class distributions before applying a precomputed scalar threshold, whereas NCP retrieves calibration neighbors and

Table 5: CFCP against NCP at matched set size. For every dataset/score cell the set size is fixed to CFCP's and NCP is moved along its own calibration curve to the same size, which is exact because the miscoverage level enters NCP only after its k-NN search. Both methods therefore spend the same set-size budget and differ only in how evenly they spend it. Values are means over ten splits with their 95% CIs. **Bold** marks the better of each pair when the two CI95 intervals do not overlap, and an unbolded pair is one the intervals do not separate.

	LAC		APS		RAPS		SAPS	
Dataset	**CFCP**	**NCP**	**CFCP**	**NCP**	**CFCP**	**NCP**	**CFCP**	**NCP**
Class coverage (COV, ↑)								
CIFAR-100	$\mathbf{0.593 \pm 0.026}$	0.452 ± 0.015	$\mathbf{0.591 \pm 0.037}$	0.385 ± 0.012	$\mathbf{0.597 \pm 0.015}$	0.552 ± 0.011	$\mathbf{0.588 \pm 0.011}$	0.566 ± 0.006
ImageNet	0.607 ± 0.010	0.591 ± 0.010	$\mathbf{0.736 \pm 0.010}$	0.584 ± 0.009	0.646 ± 0.006	0.656 ± 0.006	0.609 ± 0.018	0.609 ± 0.016
ImageNet-V2	0.510 ± 0.011	0.499 ± 0.009	$\mathbf{0.726 \pm 0.006}$	0.566 ± 0.005	0.567 ± 0.005	0.570 ± 0.004	0.510 ± 0.012	0.504 ± 0.012
WOS-46985	$\mathbf{0.654 \pm 0.021}$	0.644 ± 0.004	$\mathbf{0.667 \pm 0.010}$	0.642 ± 0.013	$\mathbf{0.667 \pm 0.014}$	0.641 ± 0.010	0.719 ± 0.010	0.723 ± 0.012
Weighted under-coverage (WUC, ↓)								
CIFAR-100	$\mathbf{0.029 \pm 0.003}$	0.042 ± 0.003	$\mathbf{0.022 \pm 0.002}$	0.051 ± 0.004	0.030 ± 0.003	0.032 ± 0.004	0.032 ± 0.003	0.029 ± 0.004
ImageNet	0.037 ± 0.002	0.039 ± 0.002	$\mathbf{0.019 \pm 0.001}$	0.044 ± 0.001	0.030 ± 0.001	0.029 ± 0.001	0.038 ± 0.002	0.037 ± 0.003
ImageNet-V2	0.105 ± 0.003	0.108 ± 0.003	$\mathbf{0.043 \pm 0.001}$	0.083 ± 0.001	0.087 ± 0.001	0.085 ± 0.001	0.106 ± 0.004	0.106 ± 0.004
WOS-46985	0.025 ± 0.002	0.026 ± 0.002	$\mathbf{0.024 \pm 0.002}$	0.030 ± 0.001	$\mathbf{0.025 \pm 0.002}$	0.031 ± 0.002	0.018 ± 0.002	0.015 ± 0.002

Table 6: CFCP against NCP at matched set size, as paired per-split differences $\Delta = \cdot_{\mathrm{CFCP}} - \cdot_{\mathrm{NCP}}$. For every dataset/score cell the set size is fixed to CFCP's and NCP is moved along its own calibration curve to the same size, which is exact because the miscoverage level enters NCP only after its k-NN search. Both methods therefore spend the same set-size budget and differ only in how evenly they spend it. Because both are read on the same ten calibration/test splits, the difference is formed within each split and the ten paired differences are then summarized as mean ± Student-t 95% CI, in percentage points. **Bold** marks a cell whose interval excludes zero in CFCP's favor; underline marks one whose interval excludes zero in NCP's favor.

Dataset	**LAC**	**APS**	**RAPS**	**SAPS**
Class coverage (ΔCOV, pp - positive favors CFCP)				
CIFAR-100	$\mathbf{+14.08 \pm 2.76}$	$\mathbf{+20.58 \pm 3.74}$	$\mathbf{+4.53 \pm 1.96}$	$+2.22 \pm 2.57$
ImageNet	$\mathbf{+1.51 \pm 0.44}$	$\mathbf{+15.26 \pm 0.64}$	$\underline{-0.95 \pm 0.65}$	$+0.02 \pm 0.62$
ImageNet-V2	$\mathbf{+1.14 \pm 0.38}$	$\mathbf{+16.01 \pm 0.21}$	-0.33 ± 0.36	$\mathbf{+0.60 \pm 0.52}$
WOS-46985	$\mathbf{+1.03 \pm 0.84}$	$\mathbf{+2.49 \pm 0.94}$	$\mathbf{+2.58 \pm 1.35}$	-0.48 ± 1.84
Weighted under-coverage (ΔWUC, pp — negative favors CFCP)				
CIFAR-100	$\mathbf{-1.24 \pm 0.39}$	$\mathbf{-2.93 \pm 0.32}$	-0.21 ± 0.38	$+0.33 \pm 0.37$
ImageNet	$\mathbf{-0.19 \pm 0.03}$	$\mathbf{-2.53 \pm 0.06}$	$+0.05 \pm 0.08$	$+0.05 \pm 0.06$
ImageNet-V2	$\mathbf{-0.26 \pm 0.05}$	$\mathbf{-3.91 \pm 0.06}$	$\underline{+0.15 \pm 0.07}$	-0.04 ± 0.06
WOS-46985	-0.07 ± 0.07	$\mathbf{-0.59 \pm 0.08}$	$\mathbf{-0.61 \pm 0.14}$	$\underline{+0.30 \pm 0.13}$

Decided cells (CFCP/NCP/undecided): COV 11/1/4; WUC 8/2/6.

constructs a localized threshold. CFCP stores centroids and their class distributions. With a flat index, centroid similarity and probability mixing cost $O(Kd + mC)$ per query, compared with $O(n_{\mathrm{cal}}d)$ for exhaustive NCP neighbor similarities. The full CFCP query includes $O(mC)$ probability mixing and the chosen score's cost in addition to centroid retrieval. Holding K, m, C, d fixed, increasing the frequency sample changes the stored statistics without increasing their dimensions.

C.3 Matching protocol

Every comparison in Section 4 places two methods at a common operating point before their class-wise behaviour is read. This subsection states what is held fixed, which data determine it and what the resulting numbers do and do not claim.

- **Coverage matching (Table 1).** All methods are calibrated at the nominal $\alpha = 0.1$ with one exception. NCP weights calibration scores by distance and takes a quantile over a query's k nearest neighbours, which forfeits exchangeability, so its realized marginal coverage is not pinned by its nominal level and reading it at $\alpha = 0.1$ would compare it at a different coverage than CFCP. Its four rows per setting are therefore re-calibrated: we sweep the

calibration level and report the fit whose realized marginal coverage is closest to that of the corresponding CFCP row, which is the matching target.

- **Selecting the comparator.** The matched tables compare CFCP against one baseline per cell, selected in two stages from the reported five baseline methods (Split, ICP, CCP, NCP, RC3P). First, the method that is best in that cell on the metric the table reports, highest COV for the COV tables, lowest WUC for the WUC tables. Second, the other methods moved to the same target takes the slot wherever it beats that pick on the same criterion. NCP holds 26 of the 28 cells, Split holds ImageNet/LAC, ImageNet/SAPS and ImageNet-V2/LAC.

- **Size matching.** Within each split, the selected baseline's calibration level is interpolated on its own α-swept curve to the mean set size CFCP produces on that split.

- **Shared size ranges.** Range-average comparisons integrate each metric against mean set size over a window on which both curves are supported on every split: the largest per-curve minimum size and the smallest per-curve maximum, taken over all (method, split) curves in the cell. On the long-tailed settings, where five baselines are swept, the window is formed per CFCP-baseline pair instead, since an intersection over all six curves would be set by the narrowest and discard most of each comparison. A consequence is that CFCP's area differs between pairs and areas are comparable only within a pair. Areas are divided by the window width and reported in percentage points, so they read as range averages comparable to the matched-point tables. Sweep levels are placed by inverting a reference size(α) at knots spread geometrically in set size rather than by stepping α, because size(α) is strongly non-linear and uniform steps in α undersample the steep end. NCP's level is floored at 0.02, below which its (λ_L, k) re-selection makes size(α) jump rather than rise.

- **Retrospective evaluation, not an operating-point rule.** All three matching steps read test-split quantities: realized marginal coverage, realized mean set size, and the range of sizes the swept curves cover. They are a measurement device for removing the size confound from a comparison, applied symmetrically to both arms, and not a procedure a practitioner could execute at deployment time, where the realized coverage and size of a candidate level are unknown. Nothing in the validity claim depends on them: Theorem 1 fixes α before the final calibration sample is used, as Appendix B requires, and each reported fit is an ordinary split-conformal predictor calibrated at its own level, valid at that level. What the matched numbers estimate is the classwise behaviour of each method conditional on being operated at a common coverage or a common size budget. Which level attains that budget on future data is a separate question this evaluation does not answer. The range-average comparisons exist for the same reason, removing the choice of a single matched point by integrating over the interval where both methods are supported. A deployable alternative, which we leave to future work, would fix the calibration level from tuning data alone: estimate the expected deployed set size at each achievable threshold (the k-th order statistic of n_q calibration scores) and select the largest $k \geq k_\alpha$ within a budget. Because k is fixed before D_q is used, Theorem 1 would still apply.

Table 7: Paired per-split shift in the strongest baseline's marginal coverage between two set sizes on its own calibration curve, $\Delta\mathrm{Cov} = \mathrm{Cov}_B(s_{\mathrm{CFCP}}) - \mathrm{Cov}_B(s_{\mathrm{main}})$, in percentage points. s_{CFCP} is the mean set size CFCP produces on that split, the matched point of Table 9, s_{main} is the baseline's mean set size at its operating point in Table 1. The difference is formed within each of the ten calibration/test splits and summarized as mean $\pm$ Student-t 95% CI. Negative values mean the baseline loses coverage when held to CFCP's set-size budget. **Size** is s_{CFCP} with s_{main} in parentheses. Underline marks an interval that excludes zero below it, **bold** one that excludes zero above it.

	LAC		APS		RAPS		SAPS	
Dataset	**Size**	ΔCov (pp)	**Size**	ΔCov (pp)	**Size**	ΔCov (pp)	**Size**	ΔCov (pp)
CIFAR-100	6.62 (8.88)	$\underline{-2.37 \pm 0.14}$	8.11 (13.74)	$\underline{-4.59 \pm 0.30}$	7.88 (8.30)	$\underline{-0.56 \pm 0.30}$	7.57 (7.77)	$\underline{-0.23 \pm 0.11}$
CIFAR-100-LT	24.90 (50.48)	$\underline{-13.92 \pm 0.73}$	25.29 (52.00)	$\underline{-15.94 \pm 0.84}$	28.82 (49.63)	$\underline{-9.86 \pm 0.76}$	28.90 (49.36)	$\underline{-9.89 \pm 0.77}$
ImageNet	2.56 (2.65)	$\underline{-0.29 \pm 0.28}$	13.29 (30.07)	$\underline{-3.80 \pm 0.12}$	3.74 (3.67)	$+0.17 \pm 0.19$	3.02 (3.02)	-0.08 ± 0.38
ImageNet-V2	3.08 (3.21)	$\underline{-0.42 \pm 0.40}$	21.42 (48.19)	$\underline{-5.50 \pm 0.21}$	4.55 (4.50)	$+0.13 \pm 0.15$	3.32 (3.38)	-0.34 ± 0.61
ImageNet-LT	5.88 (7.12)	$\underline{-1.20 \pm 0.32}$	25.61 (67.40)	$\underline{-5.97 \pm 0.19}$	6.56 (8.60)	$\underline{-1.79 \pm 0.29}$	6.54 (7.82)	$\underline{-1.10 \pm 0.41}$
ImageNet-LT-V2	6.58 (7.52)	$\underline{-1.14 \pm 0.42}$	37.60 (84.06)	$\underline{-6.48 \pm 0.32}$	7.02 (8.07)	$\underline{-1.10 \pm 0.29}$	6.88 (7.59)	$\underline{-0.92 \pm 0.41}$
WOS-46985	1.31 (1.32)	-0.04 ± 0.10	1.73 (1.99)	$\underline{-0.84 \pm 0.13}$	1.69 (1.89)	$\underline{-0.79 \pm 0.10}$	1.94 (1.56)	$\mathbf{+0.86 \pm 0.21}$

Baseline: CIFAR-100: NCP (all). CIFAR-100-LT: NCP (all). ImageNet: Split (LAC), NCP (APS), NCP (RAPS), NCP (SAPS). ImageNet-V2: NCP (LAC), NCP (APS), NCP (RAPS), Split (SAPS). ImageNet-LT: NCP (all). ImageNet-LT-V2: NCP (all). WOS-46985: NCP (all). Decided cells (lower/higher/undecided): 22/1/5.

C.4 Additional results

Table 8: Paired per-split difference in weighted under-coverage, $\Delta\mathrm{WUC} = \mathrm{WUC}_{\mathrm{CFCP}} - \mathrm{WUC}_B$, against the strongest baseline in each cell at matched efficiency. For every split the baseline's calibration level is interpolated to the mean set size CFCP produces there, so the two sides spend the same set-size budget; on CIFAR-100, ImageNet, ImageNet-V2 and WOS-46985 CFCP is likewise read off its own alpha-swept curve at that size, and on the long-tailed rows at its own operating point. Both methods are read on the same ten calibration/test splits, so the difference is formed within each split and the ten paired differences are then summarized as mean $\pm$ Student-t 95% CI, in percentage points. Negative values favor CFCP. **Bold** marks a cell whose interval is in CFCP's favor, underline marks one whose interval is in the baseline's favor. The curves are those of Figure 3 and Table 11; Table 10 reports the two methods separately at the same matched point but from the point-alpha sweep, a different set of fits, so the difference of its columns is not this ΔWUC.

Dataset	ΔWUC (pp)			
	LAC	**APS**	**RAPS**	**SAPS**
CIFAR-100	$\mathbf{-1.24 \pm 0.36}$	$\mathbf{-2.89 \pm 0.32}$	-0.13 ± 0.31	$\underline{+0.47 \pm 0.25}$
CIFAR-100-LT	$\mathbf{-4.00 \pm 0.56}$	$\mathbf{-8.52 \pm 0.72}$	$\mathbf{-1.71 \pm 0.54}$	$\mathbf{-1.37 \pm 0.55}$
ImageNet	$\mathbf{-0.19 \pm 0.03}$	$\mathbf{-2.53 \pm 0.06}$	$+0.05 \pm 0.08$	$+0.05 \pm 0.06$
ImageNet-V2	$\mathbf{-0.26 \pm 0.05}$	$\mathbf{-3.91 \pm 0.06}$	$\underline{+0.15 \pm 0.07}$	-0.04 ± 0.06
ImageNet-LT	$\mathbf{-1.24 \pm 0.46}$	$\mathbf{-8.89 \pm 0.40}$	$\mathbf{-1.03 \pm 0.10}$	$\mathbf{-0.77 \pm 0.07}$
ImageNet-LT-V2	$\mathbf{-1.11 \pm 0.06}$	$\mathbf{-5.46 \pm 0.13}$	$\mathbf{-0.79 \pm 0.09}$	$\mathbf{-0.66 \pm 0.11}$
WOS-46985	-0.06 ± 0.06	$\mathbf{-0.66 \pm 0.09}$	$\mathbf{-0.60 \pm 0.09}$	$\underline{+0.33 \pm 0.15}$

Matched baseline: CIFAR-100: NCP (all). CIFAR-100-LT: RC3P (LAC), ICP (APS), ICP (RAPS), ICP (SAPS). ImageNet: NCP (all). ImageNet-V2: NCP (all). ImageNet-LT: CCP (LAC), CCP (APS), NCP (RAPS), NCP (SAPS). ImageNet-LT-V2: NCP (all). WOS-46985: NCP (all). Decided cells (CFCP/baseline/undecided): 20/3/5.

Table 9: Class coverage (COV, ↑) of CFCP against the strongest baseline in each cell at matched efficiency. For every cell the baseline's calibration level is interpolated, per split, to the mean set size CFCP itself produces there, so the two sides spend the same set-size budget and differ only in how many classes they leave under-covered. Values are means over ten repeated splits with their 95% CIs. **Bold** marks the method with the higher COV in a pair when the two 95% confidence intervals over the repeated splits do not overlap. Pairs with no bold entry are undecided.

	LAC		APS		RAPS		SAPS	
Dataset	**CFCP**	**Best**	**CFCP**	**Best**	**CFCP**	**Best**	**CFCP**	**Best**
CIFAR-100	$\mathbf{0.593 \pm 0.026}$	0.452 ± 0.015	$\mathbf{0.591 \pm 0.037}$	0.385 ± 0.022	$\mathbf{0.597 \pm 0.025}$	0.552 ± 0.017	0.588 ± 0.021	0.566 ± 0.026
CIFAR-100-LT	$\mathbf{0.580 \pm 0.029}$	0.432 ± 0.015	$\mathbf{0.581 \pm 0.037}$	0.356 ± 0.025	$\mathbf{0.583 \pm 0.030}$	0.516 ± 0.015	$\mathbf{0.585 \pm 0.032}$	0.519 ± 0.013
ImageNet	0.607 ± 0.010	0.593 ± 0.009	$\mathbf{0.736 \pm 0.010}$	0.584 ± 0.009	0.646 ± 0.006	0.656 ± 0.006	0.609 ± 0.018	0.609 ± 0.016
ImageNet-V2	0.510 ± 0.011	0.499 ± 0.009	$\mathbf{0.726 \pm 0.006}$	0.566 ± 0.005	0.567 ± 0.005	0.570 ± 0.004	0.510 ± 0.012	0.504 ± 0.016
ImageNet-LT	0.666 ± 0.010	0.655 ± 0.010	$\mathbf{0.741 \pm 0.012}$	0.559 ± 0.005	$\mathbf{0.682 \pm 0.012}$	0.649 ± 0.013	0.681 ± 0.013	0.662 ± 0.013
ImageNet-LT-V2	0.566 ± 0.008	0.567 ± 0.010	$\mathbf{0.723 \pm 0.010}$	0.570 ± 0.007	$\mathbf{0.583 \pm 0.007}$	0.564 ± 0.007	0.573 ± 0.008	0.564 ± 0.011
WOS-46985	0.654 ± 0.021	0.644 ± 0.019	0.667 ± 0.020	0.642 ± 0.013	0.667 ± 0.020	0.641 ± 0.018	0.719 ± 0.020	0.723 ± 0.032

Matched baseline: CIFAR-100: NCP (all). CIFAR-100-LT: NCP (all). ImageNet: Split (LAC), NCP (APS), NCP (RAPS), NCP (SAPS). ImageNet-V2: NCP (LAC), NCP (APS), NCP (RAPS), Split (SAPS). ImageNet-LT: NCP (all). ImageNet-LT-V2: NCP (all). WOS-46985: NCP (all).

Table 10: Weighted under-coverage (WUC, ↓) of CFCP against the strongest baseline in each cell at matched efficiency. For every cell the baseline's calibration level is interpolated, per split, to the mean set size CFCP itself produces there, so the two sides spend the same set-size budget and differ only in how evenly they spend it. Values are means over ten repeated splits with their 95% CIs. **Bold** marks the method with the lower WUC in a pair when the two 95% confidence intervals over the repeated splits do not overlap. Pairs with no bold entry are undecided.

	LAC		APS		RAPS		SAPS	
Dataset	**CFCP**	**Best**	**CFCP**	**Best**	**CFCP**	**Best**	**CFCP**	**Best**
CIFAR-100	$\mathbf{0.029 \pm 0.003}$	0.042 ± 0.003	$\mathbf{0.022 \pm 0.002}$	0.051 ± 0.004	0.030 ± 0.003	0.032 ± 0.004	0.032 ± 0.003	0.029 ± 0.004
CIFAR-100-LT	$\mathbf{0.032 \pm 0.004}$	0.072 ± 0.005	$\mathbf{0.031 \pm 0.004}$	0.116 ± 0.008	$\mathbf{0.036 \pm 0.004}$	0.053 ± 0.006	$\mathbf{0.036 \pm 0.004}$	0.050 ± 0.006
ImageNet	0.037 ± 0.002	0.039 ± 0.002	$\mathbf{0.019 \pm 0.001}$	0.044 ± 0.001	0.030 ± 0.001	0.029 ± 0.001	0.038 ± 0.002	0.037 ± 0.003
ImageNet-V2	0.105 ± 0.003	0.108 ± 0.003	$\mathbf{0.043 \pm 0.001}$	0.083 ± 0.001	0.087 ± 0.001	0.085 ± 0.001	0.106 ± 0.004	0.106 ± 0.004
ImageNet-LT	$\mathbf{0.044 \pm 0.002}$	0.057 ± 0.004	$\mathbf{0.026 \pm 0.001}$	0.115 ± 0.005	$\mathbf{0.047 \pm 0.002}$	0.057 ± 0.002	$\mathbf{0.047 \pm 0.002}$	0.055 ± 0.003
ImageNet-LT-V2	$\mathbf{0.107 \pm 0.003}$	0.118 ± 0.003	$\mathbf{0.051 \pm 0.002}$	0.106 ± 0.003	$\mathbf{0.108 \pm 0.002}$	0.116 ± 0.002	$\mathbf{0.110 \pm 0.003}$	0.116 ± 0.004
WOS-46985	0.025 ± 0.002	0.026 ± 0.002	$\mathbf{0.024 \pm 0.002}$	0.030 ± 0.001	$\mathbf{0.025 \pm 0.002}$	0.031 ± 0.002	0.018 ± 0.002	0.016 ± 0.002

Matched baseline: CIFAR-100: NCP (all). CIFAR-100-LT: RC3P (LAC), ICP (APS), ICP (RAPS), ICP (SAPS). ImageNet: NCP (all). ImageNet-V2: NCP (all). ImageNet-LT: CCP (LAC), CCP (APS), NCP (RAPS), NCP (SAPS). ImageNet-LT-V2: NCP (all). WOS-46985: NCP (all).

Table 11: Paired per-split range-average difference in WUC, i.e. the paired difference in area under the WUC vs. mean-set-size curve divided by the width of the cell's set-size window, in percentage points. Each cell integrates both methods over the same set-size window, the range every method's own curve covers on every split, so the comparison holds across the whole coverage-efficiency trade-off rather than at one matched operating point. Both methods are read on the same ten calibration/test splits, so the difference is formed within each split and the ten paired differences are summarized as mean $\pm$ Student-t 95% CI. Negative values favor CFCP. **Bold** marks a cell whose intervals are in CFCP's favor, underline marks one whose in the baseline's favor.

Dataset	$\Delta\mathrm{AUC}_{\mathrm{WUC}}$ (pp)			
	LAC	**APS**	**RAPS**	**SAPS**
CIFAR-100	$\mathbf{-1.20 \pm 0.17}$	$\mathbf{-2.01 \pm 0.20}$	-0.11 ± 0.13	-0.09 ± 0.11
CIFAR-100-LT	$\mathbf{-2.78 \pm 0.20}$	$\mathbf{-4.99 \pm 0.15}$	$\mathbf{-2.29 \pm 0.17}$	$\mathbf{-2.06 \pm 0.20}$
ImageNet	$\mathbf{-0.14 \pm 0.01}$	$\mathbf{-1.26 \pm 0.03}$	$+0.05 \pm 0.02$	$+0.09 \pm 0.03$
ImageNet-V2	$\mathbf{-0.31 \pm 0.01}$	$\mathbf{-2.25 \pm 0.02}$	$+0.24 \pm 0.04$	$+0.21 \pm 0.06$
ImageNet-LT	$\mathbf{-0.70 \pm 0.09}$	$\mathbf{-4.41 \pm 0.22}$	$\mathbf{-0.49 \pm 0.03}$	$\mathbf{-0.44 \pm 0.03}$
ImageNet-LT-V2	$\mathbf{-1.23 \pm 0.03}$	$\mathbf{-3.21 \pm 0.02}$	$\mathbf{-0.45 \pm 0.04}$	$\mathbf{-0.40 \pm 0.04}$
WOS-46985	$+0.03 \pm 0.08$	$\mathbf{-0.38 \pm 0.04}$	$\mathbf{-0.18 \pm 0.05}$	$+0.24 \pm 0.03$

Matched baseline: CIFAR-100: NCP (all). CIFAR-100-LT: RC3P (LAC), ICP (APS), ICP (RAPS), ICP (SAPS). ImageNet: NCP (all). ImageNet-V2: NCP (all). ImageNet-LT: CCP (LAC), CCP (APS), NCP (RAPS), NCP (SAPS). ImageNet-LT-V2: NCP (all). WOS-46985: NCP (all). Decided cells (CFCP/baseline/undecided): 20/5/3.

C.5 Ablation details

Table 12: Ablation of CFCP's two components on CIFAR-100 at $\alpha = 0.1$, five splits. $r \equiv 0$ is the base-model conformal predictor. $r \equiv 1$ is the unshrunk cluster mixture. $r(x)$ is CFCP as defined. **Bold** marks the smallest set size in each row. Confidence intervals are in Table 14.

	$r \equiv 0$ (π_x)		$r \equiv 1$ (no shrink.)		$r(x)$ (CFCP)	
Score	Coverage	Size	Coverage	Size	Coverage	Size
LAC	0.894	7.97	0.899	6.27	0.897	**6.24**
APS	0.895	11.20	0.926	11.00	0.912	**8.46**
RAPS	0.897	8.32	0.901	**7.01**	0.899	7.21
SAPS	0.897	8.20	0.900	7.50	0.899	**7.36**

Protocol. Both ablations reuse the published CFCP configuration of each dataset and change only the reliability weight r. Centroids, cluster frequencies, priors and the Dirichlet smoothing are identical across arms, so the three arms differ solely in the probability matrix handed to the set constructor. Each arm receives its own split-conformal threshold, computed from its own calibration scores through the same score function, so all arms are compared at a common nominal level $\alpha = 0.1$ rather than at a common threshold. No model forward passes are involved: the ablation operates on the cached embeddings, logits and centroids of the main experiments. Reported quantities are means over five repeated calibration/test splits, $\pm$ the half-width of the 95% Student-t interval. We count a dominance for CFCP in a split when it attains at least the coverage of the ablated arm and a set size no larger, which makes the comparison insensitive to the residual coverage differences between arms.

Ablation 1: $r \equiv 0$ (no cluster frequencies). Table 14 reports every cell. On CIFAR-100 the full method dominates π_x in 4-5 of 5 splits for all four constructors, with coverage higher in the mean for every constructor and set size 10.2%-24.4% smaller. On WOS-46985 the picture splits by constructor: APS dominates in 5/5 splits (-18.5%), whereas LAC is a tie (-0.1%, 1/5), so the measured set-size difference is small in that cell.

Ablation 2: $r \equiv 1$ (no reliability shrinkage). The shrinkage yields a smaller mean set size than the unshrunk mixture in 5 of the 6 dataset$\times$constructor cells we tested, though the magni-

Table 13: Ablations at CFCP's operating point, paired over five splits at $\alpha = 0.1$ (mean $\pm$ 95% CI). Each ablated arm is moved along its own coverage–size curve. Δsize: relative change of CFCP's mean set size against the arm's size at CFCP's coverage (negative favors CFCP); wins: splits in which CFCP's set is smaller. Δcov: CFCP's coverage minus the arm's coverage at CFCP's set size, in percentage points (positive favors CFCP).

		CFCP	vs. $r \equiv 0$ (π_x)			vs. $r \equiv 1$ (no shrink.)		
Data	**Score**	Cov. / Size	Δsize (%)	wins	Δcov (pp)	Δsize (%)	wins	Δcov (pp)
CIFAR-100	LAC	0.897 / 6.24	$\mathbf{-23.3 \pm 6.8}$	5/5	$\mathbf{+2.27 \pm 0.96}$	$+2.2 \pm 3.6$	1/5	-0.07 ± 0.28
	APS	0.912 / 8.46	$\mathbf{-37.9 \pm 3.6}$	5/5	$\mathbf{+4.26 \pm 0.78}$	$\mathbf{-14.3 \pm 2.2}$	5/5	$\mathbf{+1.78 \pm 0.47}$
	RAPS	0.899 / 7.21	$\mathbf{-16.3 \pm 4.4}$	5/5	$\mathbf{+1.66 \pm 0.35}$	$+6.8 \pm 5.5$	0/5	-0.38 ± 0.27
	SAPS	0.899 / 7.36	$\mathbf{-12.0 \pm 2.9}$	5/5	$\mathbf{+1.12 \pm 0.52}$	-0.9 ± 4.5	3/5	$+0.00 \pm 0.39$
WOS-46985	LAC	0.902 / 1.17	-0.6 ± 0.7	3/5	$+0.06 \pm 0.08$	$\mathbf{-3.8 \pm 1.7}$	5/5	$\mathbf{+0.44 \pm 0.23}$
	APS	0.947 / 2.78	$\mathbf{-22.8 \pm 1.6}$	5/5	$\mathbf{+1.17 \pm 0.13}$	-3.9 ± 4.3	4/5	$+0.18 \pm 0.18$

Bold: the 95% CI excludes zero in CFCP's favor. CIFAR-100/RAPS against $r \equiv 1$ excludes zero in the arm's favor.

tude varies by two orders: -23.1% (CIFAR-100/APS) and -20.6% (WOS/APS) against -1.9% (CIFAR-100/SAPS), -2.5% (WOS/LAC) and -0.4% (CIFAR-100/LAC). The set is smaller in every split except in CIFAR-100/LAC (3/5). These reductions are not coverage-neutral: the unshrunk mixture covers more on average in four of the six cells, most visibly under APS (0.926 against 0.912 on CIFAR-100, 0.956 against 0.947 on WOS-46985). Part of the APS reduction therefore removes the unshrunk arm's excess coverage, and the shrinkage dominates it split-by-split only in WOS/LAC (4/5) and CIFAR-100/SAPS (3/5). The exception is CIFAR-100/RAPS, where the unshrunk mixture is smaller in every split ($+2.9\%$ for CFCP). The result demonstrates that the benefit of the final mixture depends on the score constructor. This ablation does not isolate the reason for that dependence. Because the arms sit at different coverage, the comparison at matched coverage below is the one we use to assess $r(x)$.

Ablations at matched coverage. To separate set-size gains from coverage differences, we keep CFCP calibrated at $\alpha = 0.1$ and move each ablated arm along its own coverage-size curve to CFCP's operating point, within each split. The curve is evaluated exactly at every calibration order statistic, i.e. at every threshold split conformal can produce for that arm, each a valid predictor at some level α_k, and read by linear interpolation between the two bracketing thresholds (adjacent thresholds differ by at most 0.002 in coverage). We report the ablated arm's set size at CFCP's realized coverage and, conversely, its coverage at CFCP's mean set size. As in the coverage-matched NCP comparison, the operating point is located on the test split, so this is a paired comparison of operating points rather than a deployable calibration rule. Table 13 reports every cell. Both readings give the same verdict in all twelve comparisons.

Against π_x, removing the coverage difference strengthens the result: CFCP's sets are smaller in every split of every CIFAR-100 cell and of WOS-46985/APS, by 12.0% to 37.9%, and WOS-46985/LAC is a tie. Against $r \equiv 1$, the reliability weight is smaller in every split in CIFAR-100/APS (-14.3%) and WOS-46985/LAC (-3.8%). CIFAR-100/LAC, CIFAR-100/SAPS and WOS-46985/APS are ties, and CIFAR-100/RAPS favors the unshrunk mixture ($+6.8\%$, 0/5). The nominal-α differences in Table 14 therefore overstate the APS gains of $r(x)$: at matched coverage the CIFAR-100/APS reduction is 14.3% rather than 23.1%, and the WOS-46985/APS reduction of 20.6% is within noise. The one cell in which the weight gains at matched coverage relative to nominal α is WOS-46985/LAC, where the cluster mixture is a worse probability estimate than π_x. This is consistent with $r(x)$ acting as a fallback when the cluster evidence is unreliable, as the NLL results below also indicate.

Ablation 2, independent of the constructor. Because r enters before the set constructor, its effect on q_x can be measured without any conformal wrapper. On CIFAR-100 the shrunk mixture is the best of the three probability vectors: the mean per-sample gain in negative log-likelihood over π_x is $+0.747 \pm 0.058$ for $r(x)$ against $+0.584 \pm 0.070$ for $r \equiv 1$, i.e. the interpolation beats both endpoints. On WOS-46985 the mixture is worse than π_x (-0.209 ± 0.006 unshrunk), and the shrinkage recovers most of that loss (-0.033 ± 0.004): on that dataset the final vector still has worse NLL than π_x. The

Table 14: Full ablation results, mean ± 95% CI over five splits at $\alpha = 0.1$. "Dominant" counts splits in which CFCP covers at least as much as the ablated arm with a set size no larger. Δsize is the mean relative change of CFCP against that arm. A smaller set without dominance means the ablated arm covered more in that split.

		$r \equiv 0$ (π_x)		$r \equiv 1$ (no shrink.)		$r(x)$ (CFCP)		Dominant / Δsize	
Data	**Score**	Coverage	Size	Coverage	Size	Coverage	Size	$r \equiv 0$	$r \equiv 1$
CIFAR-100	LAC	0.894 ± 0.007	7.97 ± 0.19	0.899 ± 0.010	6.27 ± 0.20	0.897 ± 0.010	6.24 ± 0.24	4/5, -21.6%	0/5, -0.4%
	APS	0.895 ± 0.009	11.20 ± 0.31	0.926 ± 0.006	11.00 ± 0.22	0.912 ± 0.004	8.46 ± 0.21	5/5, -24.4%	0/5, -23.1%
	RAPS	0.897 ± 0.006	8.32 ± 0.14	0.901 ± 0.007	7.01 ± 0.20	0.899 ± 0.004	7.21 ± 0.08	4/5, -13.4%	0/5, $+2.9\%$
	SAPS	0.897 ± 0.007	8.20 ± 0.17	0.900 ± 0.008	7.50 ± 0.21	0.899 ± 0.004	7.36 ± 0.10	4/5, -10.2%	3/5, -1.9%
WOS-46985	LAC	0.901 ± 0.004	1.172 ± 0.008	0.900 ± 0.005	1.200 ± 0.015	0.902 ± 0.005	1.170 ± 0.015	1/5, -0.1%	4/5, -2.5%
	APS	0.944 ± 0.004	3.42 ± 0.18	0.956 ± 0.004	3.51 ± 0.10	0.947 ± 0.004	2.79 ± 0.12	5/5, -18.5%	0/5, -20.6%

NLL and conformal results therefore measure different aspects of performance. Neither is used as a guarantee about the other.

Cluster-level coverage dispersion. Assigning every test point softly to clusters with the same weights q_x uses, and weighting each cluster by its test mass, gives the standard deviation of per-cluster coverage:

Dataset	**Score**	$r \equiv 0$	$r \equiv 1$	$r(x)$
CIFAR-100	LAC	0.0654	**0.0647**	0.0657
	APS	0.0637	0.0435	**0.0386**
	RAPS	0.0743	**0.0662**	0.0690
	SAPS	0.0740	**0.0699**	0.0719
WOS-46985	LAC	0.0920	0.0965	**0.0918**
	APS	0.0436	**0.0317**	0.0374

For APS on CIFAR-100, the full mixture reduces the reported dispersion to 0.0386 from 0.0637 for π_x. Across the remaining cells, the effect depends on the dataset, score, and mixture component. These are descriptive measurements of soft cluster-weighted coverage, not coverage guarantees for individual input regions. In particular, low dispersion alone does not ensure that every region attains the target. We do not claim general per-cluster coverage homogenization.

Limitations of these ablations. The arms sit at slightly different realized coverage at nominal α (most visibly $r \equiv 1$ under APS on CIFAR-100, 0.926 against 0.895 for π_x), so the raw set-size differences in Table 14 should be read together with realized coverage. Table 13 removes this difference, but locates the matched operating point on the test split. The dominance counts record favorable joint outcomes on the evaluated splits, without establishing population dominance. We report both. The WOS-46985 APS rows are evaluated at $\alpha = 0.1$ rather than at that cell's calibrated level, and are therefore more conservative than the corresponding main-table entry. Finally, $r(x)$ is built from assignment confidence and cluster support, and in these same runs it tracks cluster geometry closely (Spearman $+0.79$ with per-cluster compactness on CIFAR-100) but not where the cluster evidence empirically helps most (Spearman -0.20 with the per-cluster relative gain in negative log-likelihood), which is the clearest direction we see for improving the component.

C.6 Failure-mode case study on CIFAR-100

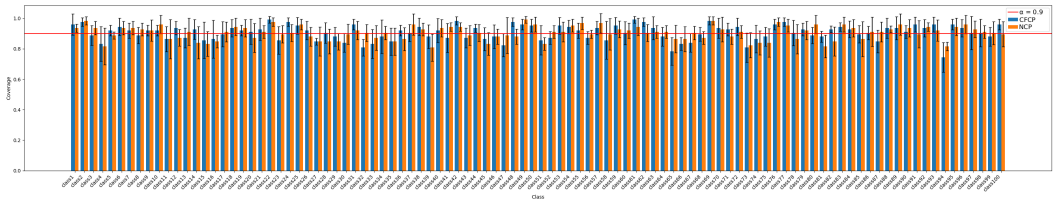
To better understand residual tail failures, we compare CFCP against the closest local-neighborhood baseline, NCP, on CIFAR-100 using APS and RAPS. For each score family, we identify the worst-covered CFCP classes and report their empirical class-conditional coverage, coverage deficit relative to the nominal target $1-\alpha = 0.9$, class frequency and average prediction-set size. Since CIFAR-100 is class-balanced, all reported classes have the same empirical frequency. Class rarity therefore does not explain differences among these classes. Representation ambiguity, frequency pooling, and the scoring/calibration choices are possible contributors. This case study does not isolate their effects.

Table 15 shows that CFCP can improve the broad coverage-efficiency tradeoff while still leaving a small number of classes with larger residual under-coverage. The clearest failure case is class94,

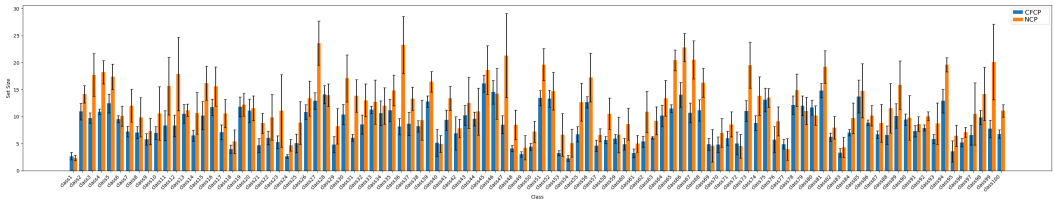
Table 15: CIFAR-100 failure-mode case study comparing CFCP and NCP at $\alpha = 0.1$. Rows show the worst-covered CFCP classes for APS and RAPS, using mean classwise coverage over ten splits. $\Delta_{0.9} = [0.9 - \widehat{\mathrm{Cov}}_c]_+$ denotes the classwise under-coverage deficit. Class names follow the standard CIFAR-100 class ordering.

Score	Class	Freq.	Coverage		$\Delta_{0.9}$		Set size	
			CFCP	NCP	CFCP	NCP	CFCP	NCP
APS	class94 (turtle)	1.0%	0.744	0.816	0.156	0.084	12.90	19.62
	class65 (possum)	1.0%	0.784	0.864	0.116	0.036	11.50	20.45
	class32 (elephant)	1.0%	0.808	0.904	0.092	0.000	8.56	13.03
RAPS	class94 (turtle)	1.0%	0.736	0.808	0.164	0.092	9.46	11.56
	class66 (rabbit)	1.0%	0.768	0.872	0.132	0.028	9.50	11.97
	class65 (possum)	1.0%	0.768	0.776	0.132	0.124	9.38	12.49

which is the worst-covered CFCP class under both APS and RAPS. For APS, CFCP attains substantially smaller sets than NCP for the same classes, but this efficiency is accompanied by a larger worst-class deficit. For RAPS, the same class remains the largest residual failure, although the difference in set size is smaller. Both methods produce relatively large sets for these classes, yet CFCP has lower empirical coverage in several of the reported comparisons. Larger sets on difficult classes do not by themselves ensure that those classes attain the target. These results motivate evaluating class or cluster-aware threshold adjustments. They also delimit the empirical contribution: improvements in aggregate COV or mean set size can coexist with worse coverage for particular classes.

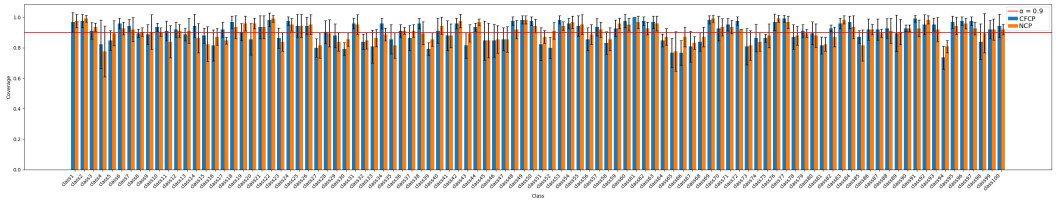


(a) Class-conditional coverage.

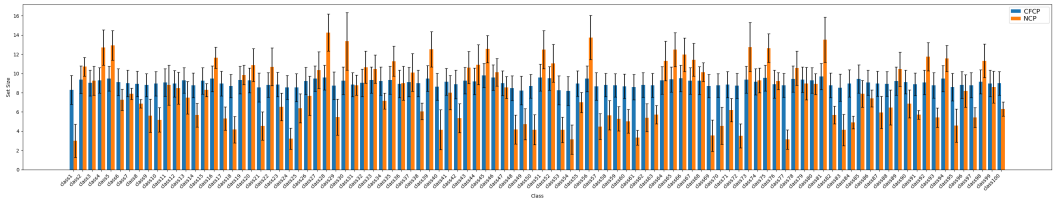


(b) Class-conditional average prediction-set size.

Figure 4: CIFAR-100 classwise comparison of CFCP-APS and NCP-APS at $\alpha = 0.1$. The red horizontal line marks the nominal target coverage $1 - \alpha = 0.9$. CFCP is generally more efficient, but the failure-mode table shows that a few classes retain larger under-coverage deficits.



(a) Class-conditional coverage.



(b) Class-conditional average prediction-set size.

Figure 5: CIFAR-100 classwise comparison of CFCP-RAPS and NCP-RAPS at $\alpha = 0.1$. CFCP covers more classes in the mean classwise view, but the worst-covered CFCP class still has a larger residual under-coverage deficit than the corresponding NCP value.