A Mathematical Notations

Table 3: **Key notations used in this paper.**

Notation	Meaning
$X \in \mathcal{X}$	Input example
$Y \in \mathcal{Y}$	The ground-truth label
f	The soft classifier
Δ_+^K	The K-dimensional probability simplex
$f(X)_y$	The predicted confidence on class y
ϵ_y^k	The class-wise top-k error for class y from f
$r_f(X, Y)$	The rank of Y predicted by $f(X)$
$\mathcal{D}_{\text{tr}}$	Training data
$\mathcal{D}_{\text{cal}}$	Calibration data
$\mathcal{D}_{\text{test}}$	Test data
n_y	The number of calibration examples for class y
$V(X, Y)$	Non-conformity scoring function
$\mathcal{C}_{1-\alpha}(X_{\text{test}})$	Prediction set for input X_{test}
α	Target mis-coverage rate
$\widehat{\alpha}_y$	Nominal mis-coverage rate for class y

B Technical Proofs of Theoretical Results

B.1 Proof of Theorem 4.1

Theorem B.1. *(Theorem 4.1 restated, class-conditional coverage of RC3P) Suppose that selecting* $\widehat{k}(y)$ *values result in the class-wise top-k error* $\epsilon_y^{\widehat{k}(y)}$ *for each class* $y \in \mathcal{Y}$. *For a target class-conditional coverage* $1 - \alpha$, *if we set* $\widehat{\alpha}_y$ *and* $\widehat{k}(y)$ *in RC3P (3) in the following ranges:*

$$\widehat{k}(y) \in \{k : \epsilon_y^k < \alpha\}, \quad 0 \le \widehat{\alpha}_y \le \alpha - \epsilon_y^{\widehat{k}(y)}, \tag{8}$$

then RC3P can achieve the class-conditional coverage for every $y \in \mathcal{Y}$:

$$\mathbb{P}_{(X,Y) \sim \mathcal{P}}\{Y \in \widehat{\mathcal{C}}_{1-\alpha}^{RC3P}(X) | Y = y\} \ge 1 - \alpha.$$

Proof. (of Theorem 4.1)

Let $y \in \mathcal{Y}$ denote any class label. In this proof, we omit the superscript k in the top-k error notation ϵ_y^k for simplicity.

With the lower bound of the coverage on class y (Theorem 1 in [45]), we have

$$
\begin{aligned}
1 - \widehat{\alpha} &\le \mathbb{P}\{Y_{\text{test}} \in \widehat{\mathcal{C}}_{1-\widehat{\alpha}}^{\text{CCP}}(X_{\text{test}}) | Y = y\} \\
&= \mathbb{P}\{V(X_{\text{test}}, Y_{\text{test}}) \le \widehat{Q}_{1-\widehat{\alpha}}^{\text{class}}(y) | Y = y\} \\
&= \mathbb{P}\{V(X_{\text{test}}, Y_{\text{test}}) \le \widehat{Q}_{1-\widehat{\alpha}}^{\text{class}}(y), r_f(X_{\text{test}}, Y_{\text{test}}) \le \widehat{k}(y) | Y = y\} \\
&\quad + \mathbb{P}\{V(X_{\text{test}}, Y_{\text{test}}) \le \widehat{Q}_{1-\widehat{\alpha}}^{\text{class}}(y), r_f(X_{\text{test}}, Y_{\text{test}}) > \widehat{k}(y) | Y = y\} \\
&\le \mathbb{P}\{V(X_{\text{test}}, Y_{\text{test}}) \le \widehat{Q}_{1-\widehat{\alpha}}^{\text{class}}(y), r_f(X_{\text{test}}, Y_{\text{test}}) \le \widehat{k}(y) | Y = y\} \\
&\quad + \underbrace{\mathbb{P}\{r_f(X_{\text{test}}, Y_{\text{test}}) > \widehat{k}(y) | Y = y\}}_{\le \epsilon_y^{\widehat{k}(y)}} \\
&\le \mathbb{P}\{Y_{\text{test}} \in \widehat{\mathcal{C}}_{1-\widehat{\alpha}}^{\text{RC3P}}(y) | Y = y\} + \epsilon_y^{\widehat{k}(y)}.
\end{aligned}
$$

Re-arranging the above inequality, we have

$$\mathbb{P}\{Y_{\text{test}} \in \widehat{\mathcal{C}}_{1-\widehat{\alpha}}^{\text{RC3P}}(y) | Y = y\} \ge 1 - \widehat{\alpha} - \epsilon_y^{\widehat{k}(y)} \ge 1 - \alpha,$$

where the last inequality is due to $\widehat{\alpha}_y \le \alpha - \epsilon_y^{\widehat{k}(y)}$. This implies that `RC3P` guarantees the class-conditional coverage on any class y. This completes the proof for Theorem 4.1. $\square$

14

B.2 Proof of Lemma 4.2

Theorem B.2. *(Lemma 4.2 restated, improved predictive efficiency of RC3P) Let $\widehat{\alpha}_y$ and $\widehat{k}(y)$ satisfy Theorem 4.1. If the following inequality holds for any $y \in \mathcal{Y}$:*

$$\mathbb{P}_{X_{test}}\left[V(X_{test}, y) \leq \widehat{Q}_{1-\widehat{\alpha}}^{class}(y), \ r_f(X_{test}, y) \leq \widehat{k}(y)\right] \leq \mathbb{P}_{X_{test}}\left[V(X_{test}, y) \leq \widehat{Q}_{1-\alpha}^{class}(y)\right], \quad (9)$$

then RC3P produces smaller expected prediction sets than CCP, i.e.,

$$\mathbb{E}_{X_{test}}[|\widehat{\mathcal{C}}_{1-\widehat{\alpha}}^{RC3P}(X_{test})|] \leq \mathbb{E}_{X_{test}}[|\widehat{\mathcal{C}}_{1-\alpha}^{CCP}(X_{test})|].$$

Proof. (of Lemma 4.2)

The proof idea is to reduce the cardinality of the prediction set made by RC3P to that made by CCP in expectation. Let $\sigma_y = \dfrac{\mathbb{P}_{X_{test}}\left[V(X_{test}, y) \leq \widehat{Q}_{1-\widehat{\alpha}}^{class}(y), \ r_f(X_{test}, y) \leq \widehat{k}(y)\right]}{\mathbb{P}_{X_{test}}\left[V(X_{test}, y) \leq \widehat{Q}_{1-\alpha}^{class}(y)\right]}$. According to the assumption in (9), we know that $\sigma_y \leq 1$, which will be used later.

We start with the expected prediction set size of RC3P and then derive its upper bound.

$$\mathbb{E}_{X_{test}}[|\widehat{\mathcal{C}}_{1-\widehat{\alpha}}^{RC3P}(X_{test})|] = \mathbb{E}_{X_{test}}\left[\sum_{y \in \mathcal{Y}} \mathbb{1}\left[V(X_{test}, y) \leq \widehat{Q}_{1-\widehat{\alpha}}^{class}(y), \ r_f(X_{test}, y) \leq \widehat{k}(y)\right]\right]$$

$$= \sum_{y \in \mathcal{Y}} \mathbb{E}_{X_{test}}\left[\mathbb{1}[V(X_{test}, y) \leq \widehat{Q}_{1-\widehat{\alpha}}^{class}(y), \ r_f(X_{test}, y) \leq \widehat{k}(y)]\right]$$

$$= \sum_{y \in \mathcal{Y}} \mathbb{P}_{X_{test}}\left[V(X_{test}, y) \leq \widehat{Q}_{1-\widehat{\alpha}}^{class}(y), \ r_f(X_{test}, y) \leq \widehat{k}(y)\right]$$

$$\stackrel{(a)}{=} \sum_{y \in \mathcal{Y}} \sigma_y \cdot \mathbb{P}_{X_{test}}\left[V(X_{test}, y) \leq \widehat{Q}_{1-\alpha}^{class}(y)\right] \quad (10)$$

$$\stackrel{(b)}{\leq} \sum_{y \in \mathcal{Y}} \mathbb{E}_{X_{test}}\left[\mathbb{1}[V(X_{test}, y) \leq \widehat{Q}_{1-\alpha}^{class}(y)]\right]$$

$$= \mathbb{E}_{X_{test}}\left[\sum_{y \in \mathcal{Y}} \mathbb{1}[V(X_{test}, y) \leq \widehat{Q}_{1-\alpha}^{class}(y)]\right] = \mathbb{E}_{X_{test}}[|\widehat{\mathcal{C}}_{1-\alpha}^{CCP}(X_{test})|], \quad (11)$$

where the equality (a) is due to the definitions of σ_y, and inequality (b) is due to the assumption

$$\sum_{y \in \mathcal{Y}} \sigma_y \cdot \mathbb{P}_{X_{test}}\left[V(X_{test}, y) \leq \widehat{Q}_{1-\alpha}^{class}(y)\right] \leq \sum_{y \in \mathcal{Y}} \mathbb{P}_{X_{test}}\left[V(X_{test}, y) \leq \widehat{Q}_{1-\alpha}^{class}(y)\right].$$

This shows that RC3P requires smaller prediction sets to guarantee the class-conditional coverage compared to CCP. □

B.3 Proof of Theorem 4.3

Theorem B.3. *(Theorem 4.3 restated, conditions of improved predictive efficiency for RC3P) Define $D = \mathbb{P}[r_f(X, y) \leq \widehat{k}(y)|Y \neq y]$, and $\bar{r}_f(X, y) = \lfloor \frac{r_f(X,y)+1}{2} \rfloor$. Denote $B = \mathbb{P}[f(X)_{(\bar{r}_f(X,y))} \leq \widehat{Q}_{1-\alpha}^{class}(y)|Y \neq y]$ if V is APS, or $B = \mathbb{P}[f(X)_{(\bar{r}_f(X,y))} + \lambda \leq \widehat{Q}_{1-\alpha}^{class}(y)|Y \neq y]$ if V is RAPS. If $B - D \geq \frac{p_y}{1-p_y}(\alpha - \epsilon_y^{\widehat{k}(y)})$, then $\sigma_y \leq 1$.*

Proof. (of Theorem 4.3)

Based on the different choices of scoring function, we first divide two scenarios:

(i): If $V(X, y)$ is the APS scoring function, since the APS score cumulatively sums the ordered prediction of $f(X)$: $V(X, y) = \sum_{l=1}^{r_f(X,y)} f(X)_{(l)}$, it is easy to verify that $V(X, y)$ is concave in

15

terms of l. As a result, we have

$$V(X,y) = \frac{r_f(X,y)}{r_f(X,y)} \cdot \sum_{l=1}^{r_f(X,y)} f(X)_{(l)} \leq r_f(X,y) \cdot f(X)_{(\lfloor \sum_{l=1}^{r_f(X,y)} l/r_f(X,y) \rfloor)} = r_f(X,y) \cdot f(X)_{(\bar{r}_f(X,y))},$$

where $\bar{r}_f(X,y) = \left\lfloor \frac{\sum_{l=1}^{r_f(X,y)} l}{r_f(X,y)} \right\rfloor = \lfloor (r_f(X,y)+1)/2 \rfloor$.

Now we lower bound $\mathbb{P}_X[V(X,y) \leq \widehat{Q}_{1-\alpha}^{\text{class}}(y)]$ as follows.

$$\begin{aligned}
&\mathbb{P}_X[V(X,y) \leq \widehat{Q}_{1-\alpha}^{\text{class}}(y)] \\
=& \underbrace{\mathbb{P}_{XY}[Y=y]}_{=p_y} \cdot \underbrace{\mathbb{P}_X[V(X,y) \leq \widehat{Q}_{1-\alpha}^{\text{class}}(y)|Y=y]}_{\geq 1-\alpha} + \underbrace{\mathbb{P}_{XY}[Y \neq y]}_{=1-p_y} \cdot \underbrace{\mathbb{P}_X[V(X,y) \leq \widehat{Q}_{1-\alpha}^{\text{class}}(y)|Y \neq y]}_{\geq B} \\
\geq& p_y(1-\alpha) + (1-p_y)B + p_y(1-\epsilon_y^{\widehat{k}(y)}) + (1-p_y)D - p_y(1-\epsilon_y^{\widehat{k}(y)}) - (1-p_y)D \\
\geq& \mathbb{P}_X[r_f(X,y) \leq \widehat{k}(y)] - p_y(\alpha - \epsilon_y^{\widehat{k}(y)}) + (1-p_y)(B-D). \quad (12)
\end{aligned}$$

According to the assumption $B - D \geq \frac{p_y}{1-p_y}(\alpha - \epsilon_y^{\widehat{k}(y)})$, we have

$$\mathbb{P}_X[r_f(X,y) \leq \widehat{k}(y)] \leq \mathbb{P}_X[V(X,y) \leq \widehat{Q}_{1-\alpha}^{\text{class}}(y)].$$

(ii): If $V(X,y)$ is the RAPS scoring function and $r_f(X,y) \leq k_{reg}$, then the RAPS scoring function could be rewritten as: $V(X,y) = \sum_{l=1}^{r_f(X,y)} f(X)_{(l)}$. As a result, we have:

$$\begin{aligned}
V(X,y) =& \frac{r_f(X,y)}{r_f(X,y)} \cdot \sum_{l=1}^{r_f(X,y)} f(X)_{(l)} \\
\leq& r_f(X,y) \cdot f(X)_{(\lfloor \sum_{l=1}^{r_f(X,y)} l/r_f(X,y) \rfloor)} \\
=& r_f(X,y) \cdot f(X)_{(\bar{r}_f(X,y))} \\
\leq& r_f(X,y) \cdot \left(f(X)_{(\bar{r}_f(X,y))} + \lambda \right).
\end{aligned}$$

If $r_f(X,y) > k_{reg}$, then the RAPS scoring function could be rewritten as: $V(X,y) = \sum_{l=1}^{r_f(X,y)} f(X)_{(l)} + \lambda(r_f(X,y) - k_{reg})$. As a result, we have

$$\begin{aligned}
V(X,y) =& \frac{r_f(X,y)}{r_f(X,y)} \cdot \left(\sum_{l=1}^{r_f(X,y)} f(X)_{(l)} + \lambda(r_f(X,y) - k_{reg}) \right) \\
\leq& r_f(X,y) \cdot \left(f(X)_{(\bar{r}_f(X,y))} + \lambda\left(1 - \frac{k_{reg}}{r_f(X,y)}\right) \right) \\
\leq& r_f(X,y) \cdot \left(f(X)_{(\bar{r}_f(X,y))} + \lambda \right).
\end{aligned}$$

Then, by applying the Inequality 12, we have:

$$\mathbb{P}_X[r_f(X,y) \leq \widehat{k}(y)] \leq \mathbb{P}_X[V(X,y) \leq \widehat{Q}_{1-\alpha}^{\text{class}}(y)].$$

This completes the proof for Theorem 4.3. $\qquad\square$

C   Complete Experimental Results

C.1   Training Details

For CIFAR-10 and CIFAR-100, we train ResNet20 using LDAM loss function given in [10] with standard mini-batch stochastic gradient descent (SGD) using learning rate 0.1, momentum 0.9, and weight decay $2e-4$ for 200 epochs and 50 epochs. The batch size is 128. For experiments on

16

mini-ImageNet, we use the same setting. For Food-101, the batch size is 256 and other parameters are kept the same. We reported our main results when models were trained in 200 epochs. Other results are reported in Appendix C.8 and Table 11.

We also evaluate the top-1 accuracy over the majority, medium, and minority groups of classes as the class-wise performance when 200 epochs. To show the variation of class-wise performance, we divide some classes with the largest number of data samples into the majority group, and the number of these classes is a quarter (25%) of the total number of classes. Similarly, we divide the classes with the smallest number of data into the minority group (25%) and the remaining classes as the medium group (50%). In the above table, we show the accuracy of three groups with three imbalance types and two imbalance ratios $\rho = 0.1, \rho = 0.5$ on four datasets.

The results are summarized in Table 4. As can be seen, the group-wise performance can vary significantly from high to very low. The class-imbalance setting is the case where the classifier does not perform very well in some classes.

Table 4: Top-1 accuracy of minority, medium, and majority groups with three imbalance types and two imbalance ratios $\rho = 0.1, \rho = 0.5$ on four datasets. We could observe that the class-wise performance varies significantly over different classes.

Groups	EXP		POLY		MAJ	
	$\rho = 0.5$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.1$
CIFAR-10						
Minority	0.913	0.961	0.932	0.901	0.940	0.927
Medium	0.872	0.822	0.867	0.847	0.848	0.75
Majority	0.949	0.832	0.933	0.948	0.914	0.795
CIFAR-100						
Minority	0.554	0.295	0.468	0.352	0.572	0.365
Medium	0.589	0.536	0.517	0.413	0.574	0.476
Majority	0.668	0.720	0.671	0.588	0.616	0.562
mini-ImageNet						
Minority	0.677	0.640	0.624	0.627	0.626	0.642
Medium	0.527	0.546	0.533	0.530	0.526	0.538
Majority	0.633	0.679	0.684	0.67	0.673	0.686
Food-101						
Minority	0.453	0.231	0.379	0.289	0.505	0.333
Medium	0.579	0.474	0.496	0.398	0.579	0.467
Majority	0.582	0.660	0.596	0.563	0.532	0.490

C.2 Calibration Details

As mentioned in Section 5.1, we balanced split the validation set of CIFAR-10 and CIFAR-100, the number of calibration data is 5000. For mini-ImageNet, the number of calibration data is 15000. For Food-101, the total number is 12625. To compute the mean and standard deviation for the overall performance, we repeat calibration experiments for 10 times. In our main results, We set $\alpha = 0.1$. We also report other experiment results of different α values, $\alpha = 0.05$ and $\alpha = 0.01$, in Appendix C.7, and Table 9 and 10.

The regularization parameter for RAPS scoring function is from the set $k_{reg} \in \{3, 5, 7\}$ and $\lambda \in \{0.001, 0.01, 0.1\}$ based on the empirical setting in `cluster-CP`. We select the combination of k_{reg} and λ for each experiment with the same imbalanced type and imbalanced ratio on the same dataset, where most of the APSS values of all methods are minimum.

The hyper-parameter g is selected from the set $\{0.25, 0.5, 0.75, 1.0\}$ to find the minimal g that `CCP`, `Cluster-CP` [3], and `RC3P` achieve the target class-conditional coverage. We clarify that for each dataset and each class-conditional CP method, we use fixed g values. The detailed g values

---

[3]https://github.com/tiffanyding/class-conditional-conformal/tree/main

17

are displayed in Table 5. From Table 5, we could observe that the hyperparameter g for `RC3P` is always smaller than other methods, which means that comparing other class-wise CP algorithms, our algorithm needs the smallest inflation on $1 - \widehat{\alpha}$ to achieve the target class-conditional coverage. This could also match the result of histograms of class-conditional coverage.

Table 5: Hyperparameter g choices for each class-conditional CP methods `CCP`, `Cluster-CP`, and `RC3P` on four datasets CIFAR-10, CIFAR-100, mini-ImageNet, and Food101. We could observe that all g values are in constant order to make a fair comparison. Meanwhile, the hyperparameter g for `RC3P` is always smaller than other methods.

Methods	Dataset			
	CIFAR-10	CIFAR-100	mini-ImageNet	FOOD-101
CCP	0.5	0.5	0.75	0.75
Cluster-CP	1.0	0.5	0.75	0.75
RC3P	0.5	0.25	0.5	0.5

C.3 Illustration of Imbalanced Data
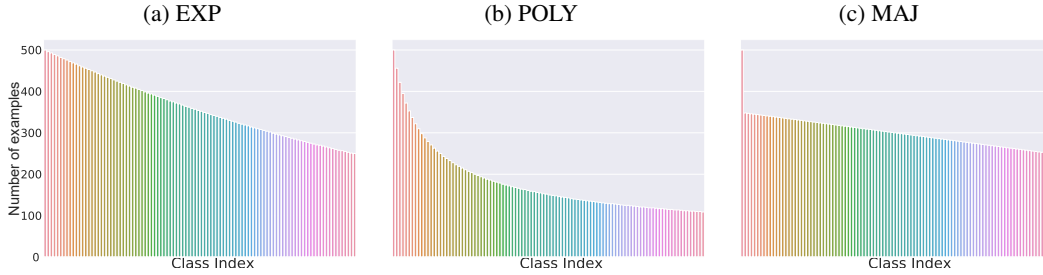
Figure 4: Illustrative examples of the different imbalanced distributions of the number of training examples per class index c on CIFAR-100

C.4 Comparison Experiments Using APS Score Function

Based on the results in Table 6, we make the following observations: (i) `CCP`, `Cluster-CP`, and `RC3P` can guarantee the class-conditional coverage; and (ii) `RC3P` significantly outperforms `CCP` and `Cluster-CCP` on three datasets by producing smaller prediction sets.

18

Table 6: Results comparing `CCP`, `Cluster-CP`, and `RC3P` with ResNet-20 model and APS scoring function under different imbalance ratios $\rho = 0.5$ and $\rho = 0.1$ when $\alpha = 0.1$. We set UCR of `RC3P` the same as or better than that of `CCP` and `Cluster-CP` for a fair comparison of prediction set size. The APSS results show that `RC3P` significantly outperforms `Cluster-CP` in terms of the average prediction set size over all settings on CIFAR-100, mini-ImageNet, and Food-101.

Measure	Methods	EXP		POLY		MAJ	
		$\rho = 0.5$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.1$
		CIFAR-10					
UCR	CCP	0.050 ± 0.016	0.100 ± 0.020	0.100 ± 0.032	$\mathbf{0.050 \pm 0.021}$	0.070 ± 0.014	$\mathbf{0.040 \pm 0.015}$
	Cluster-CP	$\mathbf{0.010 \pm 0.009}$	$\mathbf{0.090 \pm 0.009}$	$\mathbf{0.080 \pm 0.019}$	0.060 ± 0.001	$\mathbf{0.020 \pm 0.012}$	0.070 ± 0.014
	RC3P	0.050 ± 0.016	0.100 ± 0.020	0.100 ± 0.032	0.050 ± 0.021	0.070 ± 0.014	$\mathbf{0.040 \pm 0.015}$
APSS	CCP	$\mathbf{1.555 \pm 0.010}$	$\mathbf{1.855 \pm 0.014}$	$\mathbf{1.538 \pm 0.010}$	$\mathbf{1.776 \pm 0.012}$	$\mathbf{1.840 \pm 0.020}$	$\mathbf{2.629 \pm 0.013}$
	Cluster-CP	1.714 ± 0.018	2.162 ± 0.015	1.706 ± 0.014	1.928 ± 0.013	1.948 ± 0.023	3.220 ± 0.020
	RC3P	$\mathbf{1.555 \pm 0.010}$	$\mathbf{1.855 \pm 0.014}$	$\mathbf{1.538 \pm 0.010}$	$\mathbf{1.776 \pm 0.012}$	$\mathbf{1.840 \pm 0.020}$	$\mathbf{2.629 \pm 0.013}$
		CIFAR-100					
UCR	CCP	0.007 ± 0.002	$\mathbf{0.010 \pm 0.002}$	0.010 ± 0.002	$\mathbf{0.014 \pm 0.003}$	0.016 ± 0.003	$\mathbf{0.008 \pm 0.004}$
	Cluster-CP	0.012 ± 0.002	0.016 ± 0.004	0.020 ± 0.003	0.004 ± 0.002	0.016 ± 0.003	0.019 ± 0.005
	RC3P	$\mathbf{0.005 \pm 0.002}$	0.011 ± 0.002	$\mathbf{0.009 \pm 0.003}$	0.015 ± 0.003	$\mathbf{0.008 \pm 0.002}$	$\mathbf{0.008 \pm 0.004}$
APSS	CCP	44.224 ± 0.341	50.969 ± 0.345	49.889 ± 0.353	64.343 ± 0.237	44.194 ± 0.514	64.642 ± 0.535
	Cluster-CP	29.238 ± 0.609	37.592 ± 0.857	38.252 ± 0.353	52.391 ± 0.595	31.518 ± 0.335	50.883 ± 0.673
	RC3P	$\mathbf{17.705 \pm 0.004}$	$\mathbf{21.954 \pm 0.005}$	$\mathbf{23.048 \pm 0.008}$	$\mathbf{33.185 \pm 0.005}$	$\mathbf{18.581 \pm 0.007}$	$\mathbf{32.699 \pm 0.005}$
		mini-ImageNet					
UCR	CCP	0.008 ± 0.004	0.008 ± 0.004	0.005 ± 0.002	0.004 ± 0.001	0.010 ± 0.004	0.005 ± 0.002
	Cluster-CP	0.014 ± 0.004	0.012 ± 0.004	0.011 ± 0.003	0.014 ± 0.003	0.008 ± 0.002	0.010 ± 0.003
	RC3P	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.001 \pm 0.001}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$
APSS	CCP	26.676 ± 0.171	26.111 ± 0.194	26.626 ± 0.133	26.159 ± 0.208	27.313 ± 0.154	25.629 ± 0.207
	Cluster-CP	25.889 ± 0.301	25.253 ± 0.346	26.150 ± 0.393	25.633 ± 0.268	26.918 ± 0.241	25.348 ± 0.334
	RC3P	$\mathbf{18.129 \pm 0.003}$	$\mathbf{17.082 \pm 0.002}$	$\mathbf{17.784 \pm 0.003}$	$\mathbf{17.465 \pm 0.003}$	$\mathbf{18.111 \pm 0.002}$	$\mathbf{17.167 \pm 0.004}$
		Food-101					
UCR	CCP	0.006 ± 0.002	0.006 ± 0.002	0.009 ± 0.003	0.008 ± 0.001	0.006 ± 0.001	0.008 ± 0.002
	Cluster-CP	0.003 ± 0.002	0.009 ± 0.003	0.004 ± 0.001	0.009 ± 0.002	0.011 ± 0.003	0.011 ± 0.002
	RC3P	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.001 \pm 0.001}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$
APSS	CCP	27.022 ± 0.192	30.900 ± 0.170	30.943 ± 0.119	35.912 ± 0.105	27.415 ± 0.194	36.776 ± 0.132
	Cluster-CP	28.953 ± 0.333	33.375 ± 0.377	33.079 ± 0.393	38.301 ± 0.232	30.071 ± 0.412	39.632 ± 0.342
	RC3P	$\mathbf{18.369 \pm 0.004}$	$\mathbf{21.556 \pm 0.006}$	$\mathbf{21.499 \pm 0.003}$	$\mathbf{25.853 \pm 0.004}$	$\mathbf{19.398 \pm 0.006}$	$\mathbf{26.585 \pm 0.004}$

19

C.5 Comparison Experiments Using RAPS Score Function

With the same model, evaluation metrics, and RAPS score function [1], we add the comparison experiments with CCP, and Cluster-CP on four datasets with different imbalanced types and imbalance ratio $\rho = 0.5$ and $\rho = 0.1$. The regularization parameter for RAPS scoring function is from the set $k_{reg} \in \{3, 5, 7\}$ and $\lambda \in \{0.001, 0.01, 0.1\}$. We select the combination of k_{reg} and λ for each experiment with the same imbalanced type and imbalanced ratio on the same dataset, where most of the $APSS$ values of all methods are minimum. The overall performance is summarized in Table 7. We highlight that we also select the g from the set $g \in \{0.25, 0.5, 0.75, 1.0\}$ to find the minimal g that CCP, Cluster-CP, and RC3P approximately achieves the target class conditional coverage.

Based on the results in Table 7, we make the following observations: (i) CCP, Cluster-CP, and RC3P can guarantee the class-conditional coverage; and (ii) RC3P significantly outperforms CCP and Cluster-CP on three datasets by producing smaller prediction sets.

Table 7: Results comparing CCP, Cluster-CP, and RC3P with ResNet-20 model and the RAPS scoring function under different imbalance ratios $\rho = 0.5$ and $\rho = 0.1$ when $\alpha = 0.1$. The regularization parameter for RAPS scoring function is selected from the set $[3, 5, 7]$ and $[0.001, 0.01, 0.1]$. We select the best results for each element in the table. We set UCR of RC3P the same as or better than that of CCP and Cluster-CP for a fair comparison of prediction set size. The APSS results show that RC3P significantly outperforms CCP and Cluster-CP in terms of average prediction set size over all settings on CIFAR-100, mini-ImageNet, and Food-101.

Measure	Methods	EXP		POLY		MAJ	
		$\rho = 0.5$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.1$
				CIFAR-10			
UCR	CCP	0.050 ± 0.016	$\mathbf{0.010 \pm 0.020}$	0.100 ± 0.028	$\mathbf{0.050 \pm 0.021}$	0.070 ± 0.014	$\mathbf{0.040 \pm 0.015}$
	Cluster-CP	$\mathbf{0.010 \pm 0.009}$	$\mathbf{0.010 \pm 0.010}$	$\mathbf{0.080 \pm 0.019}$	0.060 ± 0.015	$\mathbf{0.020 \pm 0.025}$	0.070 ± 0.014
	RC3P	0.050 ± 0.016	$\mathbf{0.010 \pm 0.020}$	0.100 ± 0.028	$\mathbf{0.050 \pm 0.021}$	0.070 ± 0.014	$\mathbf{0.040 \pm 0.015}$
APSS	CCP	$\mathbf{1.555 \pm 0.010}$	$\mathbf{1.855 \pm 0.014}$	$\mathbf{1.538 \pm 0.010}$	$\mathbf{1.776 \pm 0.012}$	$\mathbf{1.840 \pm 0.020}$	$\mathbf{2.632 \pm 0.012}$
	Cluster-CP	1.714 ± 0.018	2.162 ± 0.015	1.706 ± 0.014	1.929 ± 0.013	1.787 ± 0.019	2.968 ± 0.024
	RC3P	$\mathbf{1.555 \pm 0.010}$	$\mathbf{1.855 \pm 0.014}$	$\mathbf{1.538 \pm 0.010}$	$\mathbf{1.776 \pm 0.012}$	$\mathbf{1.840 \pm 0.020}$	$\mathbf{2.632 \pm 0.012}$
				CIFAR-100			
UCR	CCP	0.007 ± 0.002	$\mathbf{0.011 \pm 0.002}$	0.010 ± 0.002	$\mathbf{0.015 \pm 0.003}$	0.015 ± 0.003	0.008 ± 0.004
	Cluster-CP	0.012 ± 0.002	0.017 ± 0.004	0.019 ± 0.004	0.034 ± 0.005	$\mathbf{0.008 \pm 0.003}$	0.018 ± 0.006
	RC3P	$\mathbf{0.005 \pm 0.002}$	$\mathbf{0.011 \pm 0.002}$	$\mathbf{0.009 \pm 0.003}$	$\mathbf{0.015 \pm 0.003}$	0.015 ± 0.003	$\mathbf{0.008 \pm 0.004}$
APSS	CCP	44.250 ± 0.342	50.970 ± 0.345	49.886 ± 0.353	64.332 ± 0.236	48.343 ± 0.353	64.663 ± 0.535
	Cluster-CP	29.267 ± 0.612	37.795 ± 0.862	38.258 ± 0.320	52.374 ± 0.592	31.513 ± 0.325	50.379 ± 0.684
	RC3P	$\mathbf{17.705 \pm 0.004}$	$\mathbf{21.954 \pm 0.005}$	$\mathbf{23.048 \pm 0.008}$	$\mathbf{33.185 \pm 0.005}$	$\mathbf{18.581 \pm 0.006}$	$\mathbf{32.699 \pm 0.006}$
				mini-ImageNet			
UCR	CCP	0.008 ± 0.003	0.009 ± 0.004	0.005 ± 0.002	0.004 ± 0.002	0.009 ± 0.003	0.005 ± 0.002
	Cluster-CP	0.006 ± 0.002	0.013 ± 0.005	0.009 ± 0.003	0.016 ± 0.001	0.007 ± 0.002	0.009 ± 0.004
	RC3P	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.001 \pm 0.001}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$
APSS	CCP	26.756 ± 0.178	26.212 ± 0.199	26.689 ± 0.142	26.248 ± 0.219	27.397 ± 0.162	25.725 ± 0.214
	Cluster-CP	26.027 ± 0.325	25.415 ± 0.289	26.288 ± 0.407	25.712 ± 0.315	26.969 ± 0.305	25.532 ± 0.350
	RC3P	$\mathbf{18.129 \pm 0.003}$	$\mathbf{17.082 \pm 0.002}$	$\mathbf{17.784 \pm 0.003}$	$\mathbf{17.465 \pm 0.003}$	$\mathbf{18.111 \pm 0.002}$	$\mathbf{17.167 \pm 0.004}$
				Food-101			
UCR	CCP	0.006 ± 0.003	0.006 ± 0.002	0.009 ± 0.003	0.008 ± 0.001	0.006 ± 0.002	0.008 ± 0.002
	Cluster-CP	0.004 ± 0.003	0.012 ± 0.004	0.006 ± 0.002	0.006 ± 0.003	0.011 ± 0.003	0.014 ± 0.004
	RC3P	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.001 \pm 0.001}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$
APSS	CCP	27.022 ± 0.192	30.900 ± 0.170	30.966 ± 0.125	35.940 ± 0.111	27.439 ± 0.203	36.802 ± 0.138
	Cluster-CP	28.953 ± 0.333	33.375 ± 0.377	33.337 ± 0.409	38.499 ± 0.216	29.946 ± 0.407	39.529 ± 0.306
	RC3P	$\mathbf{18.369 \pm 0.004}$	$\mathbf{21.556 \pm 0.006}$	$\mathbf{21.499 \pm 0.003}$	$\mathbf{25.853 \pm 0.004}$	$\mathbf{19.397 \pm 0.006}$	$\mathbf{26.585 \pm 0.004}$

C.6 Comparison Experiments Using HPS Score Function

With the same model, evaluation metrics, and HPS score function [1], we add the comparison experiments with CCP, and Cluster-CP on four datasets with different imbalanced types and imbalance ratio $\rho = 0.5$ and $\rho = 0.1$. The overall performance is summarized in Table 8. We highlight that we also select the g from the set $g \in \{0.25, 0.5, 0.75, 1.0\}$ to find the minimal g that CCP, Cluster-CP, and RC3P approximately achieves the target class conditional coverage.

20

Based on the results in Table 8, we make the following observations: (i) CCP, Cluster-CP, and RC3P can guarantee the class-conditional coverage; and (ii) RC3P significantly outperforms CCP and Cluster-CP on three datasets by producing smaller prediction sets.

Table 8: Results comparing CCP, Cluster-CP, and RC3P with ResNet-20 model and the HPS scoring function under different imbalance ratios $\rho = 0.5$ and $\rho = 0.1$ when $\alpha = 0.1$. We set UCR of RC3P the same as or better than that of CCP and Cluster-CP for a fair comparison of prediction set size. RC3P significantly outperforms CCP and Cluster-CP with 20.91% (four datasets) or 27.88% (excluding CIFAR-10) reduction in APSS.

Measure	Methods	EXP		POLY		MAJ	
		$\rho = 0.5$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.1$
				CIFAR-10			
UCR	CCP	0.050 ± 0.016	$\mathbf{0.010 \pm 0.020}$	0.100 ± 0.028	$\mathbf{0.050 \pm 0.021}$	0.070 ± 0.014	$\mathbf{0.040 \pm 0.015}$
	Cluster-CP	$\mathbf{0.010 \pm 0.009}$	$\mathbf{0.010 \pm 0.010}$	$\mathbf{0.080 \pm 0.019}$	0.060 ± 0.015	$\mathbf{0.020 \pm 0.025}$	0.070 ± 0.014
	RC3P	0.050 ± 0.016	$\mathbf{0.010 \pm 0.020}$	0.100 ± 0.028	$\mathbf{0.050 \pm 0.021}$	0.070 ± 0.014	$\mathbf{0.040 \pm 0.015}$
APSS	CCP	$\mathbf{1.144 \pm 0.005}$	$\mathbf{1.324 \pm 0.007}$	$\mathbf{1.137 \pm 0.003}$	$\mathbf{1.243 \pm 0.005}$	$\mathbf{1.272 \pm 0.008}$	$\mathbf{1.936 \pm 0.010}$
	Cluster-CP	1.214 ± 0.008	1.508 ± 0.010	1.211 ± 0.004	1.354 ± 0.005	1.336 ± 0.009	2.312 ± 0.025
	RC3P	$\mathbf{1.144 \pm 0.005}$	$\mathbf{1.324 \pm 0.007}$	$\mathbf{1.137 \pm 0.003}$	$\mathbf{1.243 \pm 0.005}$	$\mathbf{1.272 \pm 0.008}$	$\mathbf{1.936 \pm 0.010}$
				CIFAR-100			
UCR	CCP	0.007 ± 0.002	$\mathbf{0.011 \pm 0.002}$	0.010 ± 0.002	$\mathbf{0.015 \pm 0.003}$	0.015 ± 0.003	0.008 ± 0.004
	Cluster-CP	0.012 ± 0.002	0.017 ± 0.004	0.019 ± 0.004	0.034 ± 0.005	$\mathbf{0.008 \pm 0.003}$	0.018 ± 0.006
	RC3P	$\mathbf{0.005 \pm 0.002}$	$\mathbf{0.011 \pm 0.002}$	$\mathbf{0.009 \pm 0.003}$	$\mathbf{0.015 \pm 0.003}$	0.015 ± 0.003	$\mathbf{0.008 \pm 0.004}$
APSS	CCP	41.351 ± 0.242	49.469 ± 0.344	48.063 ± 0.376	63.963 ± 0.277	46.125 ± 0.351	64.371 ± 0.564
	Cluster-CP	27.566 ± 0.555	35.528 ± 0.979	36.101 ± 0.565	51.333 ± 0.776	29.323 ± 0.363	50.519 ± 0.679
	RC3P	$\mathbf{20.363 \pm 0.006}$	$\mathbf{25.212 \pm 0.010}$	$\mathbf{25.908 \pm 0.007}$	$\mathbf{36.951 \pm 0.018}$	$\mathbf{21.149 \pm 0.006}$	$\mathbf{35.606 \pm 0.005}$
				mini-ImageNet			
UCR	CCP	0.008 ± 0.003	0.009 ± 0.004	0.005 ± 0.002	0.004 ± 0.002	0.009 ± 0.003	0.005 ± 0.002
	Cluster-CP	0.006 ± 0.002	0.013 ± 0.005	0.009 ± 0.003	0.016 ± 0.001	0.007 ± 0.002	0.009 ± 0.004
	RC3P	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.001 \pm 0.001}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$
APSS	CCP	24.633 ± 0.212	24.467 ± 0.149	24.379 ± 0.152	24.472 ± 0.167	25.449 ± 0.196	23.885 ± 0.159
	Cluster-CP	23.911 ± 0.322	24.023 ± 0.195	24.233 ± 0.428	23.263 ± 0.295	24.987 ± 0.319	23.323 ± 0.378
	RC3P	$\mathbf{17.830 \pm 0.104}$	$\mathbf{17.036 \pm 0.014}$	$\mathbf{17.684 \pm 0.062}$	$\mathbf{17.393 \pm 0.013}$	$\mathbf{18.024 \pm 0.049}$	$\mathbf{17.086 \pm 0.059}$
				Food-101			
UCR	CCP	0.006 ± 0.003	0.006 ± 0.002	0.009 ± 0.003	0.008 ± 0.001	0.006 ± 0.002	0.008 ± 0.002
	Cluster-CP	0.004 ± 0.003	0.012 ± 0.004	0.006 ± 0.002	0.006 ± 0.003	0.011 ± 0.003	0.014 ± 0.004
	RC3P	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.001 \pm 0.001}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.000 \pm 0.000}$
APSS	CCP	26.481 ± 0.142	30.524 ± 0.152	30.787 ± 0.099	35.657 ± 0.107	26.826 ± 0.163	36.518 ± 0.122
	Cluster-CP	29.347 ± 0.288	33.806 ± 0.513	33.407 ± 0.345	38.956 ± 0.242	29.606 ± 0.436	39.880 ± 0.318
	RC3P	$\mathbf{18.337 \pm 0.004}$	$\mathbf{21.558 \pm 0.006}$	$\mathbf{21.477 \pm 0.003}$	$\mathbf{25.853 \pm 0.005}$	$\mathbf{19.396 \pm 0.008}$	$\mathbf{26.584 \pm 0.003}$

C.7 Comparison Experiments with different α values

With the same model, evaluation metrics, and scoring functions, we add the comparison experiments with CCP, and Cluster-CP on four datasets with different imbalanced types and imbalance ratio $\rho = 0.5$ and $\rho = 0.1$ under the different α values. The overall performance is summarized in Table 9 and 10, with $\alpha = 0.05$ and $\alpha = 0.01$, respectively. We highlight that we also select the g from the set $g \in [0.15, 0.75]$ with 0.05 range to find the minimal g that CCP, Cluster-CP, and RC3P approximately achieves the target class conditional coverage.

Based on the results in Table 7, we make the following observations: (i) CCP, Cluster-CP, and RC3P can guarantee the class-conditional coverage; and (ii) RC3P significantly outperforms CCP and Cluster-CP on three datasets by producing smaller prediction sets.

21

Table 9: APSS results comparing CCP, Cluster-CP, and RC3P with ResNet-20 model under different imbalance ratio $\rho = 0.5$ and $\rho = 0.1$ where $\alpha = 0.05$. For a fair comparison of prediction set size, we set UCR of RC3P the same as or smaller (more restrictive) than that of CCP and Cluster-CP under 0.16 on CIFAR-10 and 0.03 on other datasets. The APSS results show that RC3P significantly outperforms CCP and Cluster-CP in terms of average prediction set size with 21.036% (four datasets) or 28.048% (excluding CIFAR-10) reduction in prediction size on average over min{CCP, cluster-CP}.

Conformity Score	Methods	EXP		POLY		MAJ	
		$\rho = 0.5$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.1$
		CIFAR-10					
APS	CCP	**2.861 ± 0.027**	**3.496 ± 0.037**	**2.744 ± 0.033**	**3.222 ± 0.018**	**3.269 ± 0.037**	**4.836 ± 0.035**
	Cluster-CP	3.443 ± 0.041	4.551 ± 0.049	3.309 ± 0.037	4.012 ± 0.039	4.075 ± 0.069	5.958 ± 0.070
	RC3P	**2.861 ± 0.027**	**3.496 ± 0.037**	**2.744 ± 0.033**	**3.222 ± 0.018**	**3.269 ± 0.037**	**4.836 ± 0.035**
RAPS	CCP	**2.833 ± 0.018**	**3.448 ± 0.036**	**2.774 ± 0.033**	**3.231 ± 0.021**	**3.301 ± 0.024**	**4.842 ± 0.037**
	Cluster-CP	3.430 ± 0.044	4.389 ± 0.062	3.352 ± 0.035	3.876 ± 0.034	4.044 ± 0.055	5.959 ± 0.083
	RC3P	**2.833 ± 0.018**	**3.448 ± 0.036**	**2.774 ± 0.033**	**3.231 ± 0.021**	**3.301 ± 0.024**	**4.842 ± 0.037**
		CIFAR-100					
APS	CCP	44.019 ± 0.295	51.004 ± 0.366	49.564 ± 0.315	64.314 ± 0.231	48.024 ± 0.386	64.941 ± 0.532
	Cluster-CP	39.641 ± 0.567	46.746 ± 0.147	47.654 ± 0.371	62.340 ± 0.404	37.634 ± 0.537	60.841 ± 0.391
	RC3P	**32.128 ± 0.011**	**38.769 ± 0.006**	**39.930 ± 0.008**	**53.147 ± 0.010**	**34.361 ± 0.007**	**51.498 ± 0.009**
RAPS	CCP	44.234 ± 0.341	50.950 ± 0.344	49.889 ± 0.355	64.339 ± 0.236	48.310 ± 0.353	64.628 ± 0.535
	Cluster-CP	39.212 ± 0.365	46.840 ± 0.186	49.094 ± 0.280	62.095 ± 0.278	41.596 ± 0.323	60.158 ± 0.536
	RC3P	**32.135 ± 0.010**	**38.793 ± 0.007**	**39.871 ± 0.010**	**53.169 ± 0.009**	**34.380 ± 0.007**	**51.512 ± 0.008**
		mini-ImageNet					
APS	CCP	58.527 ± 0.445	57.527 ± 0.408	60.327 ± 0.520	56.581 ± 0.438	59.360 ± 0.430	56.636 ± 0.469
	Cluster-CP	47.613 ± 0.544	46.650 ± 0.699	47.117 ± 0.930	45.360 ± 0.582	59.002 ± 0.434	56.147 ± 0.456
	RC3P	**32.046 ± 0.002**	**31.729 ± 0.003**	**31.718 ± 0.004**	**32.048 ± 0.003**	**32.909 ± 0.007**	**31.441 ± 0.004**
RAPS	CCP	58.615 ± 0.428	57.626 ± 0.394	60.173 ± 0.527	56.702 ± 0.414	59.532 ± 0.430	56.903 ± 0.460
	Cluster-CP	47.427 ± 0.588	46.767 ± 0.724	47.302 ± 1.126	45.603 ± 0.639	59.408 ± 0.482	56.779 ± 0.486
	RC3P	**32.040 ± 0.003**	**31.741 ± 0.003**	**31.752 ± 0.003**	**32.067 ± 0.002**	**32.914 ± 0.005**	**31.417 ± 0.005**
		Food-101					
APS	CCP	55.967 ± 0.464	60.374 ± 0.383	60.717 ± 0.596	65.698 ± 0.405	56.934 ± 0.446	66.654 ± 0.511
	Cluster-CP	48.699 ± 0.512	55.288 ± 0.815	54.063 ± 0.885	60.104 ± 0.608	48.894 ± 0.919	59.432 ± 0.754
	RC3P	**31.224 ± 0.004**	**35.273 ± 0.007**	**35.364 ± 0.003**	**41.109 ± 0.005**	**31.661 ± 0.005**	**39.135 ± 0.003**
RAPS	CCP	55.872 ± 0.465	60.764 ± 0.394	60.618 ± 0.579	65.681 ± 0.401	56.982 ± 0.447	66.615 ± 0.504
	Cluster-CP	48.371 ± 0.513	55.155 ± 0.775	53.813 ± 0.864	59.912 ± 0.530	49.259 ± 0.846	59.307 ± 0.648
	RC3P	**31.229 ± 0.004**	**35.283 ± 0.006**	**35.379 ± 0.003**	**41.113 ± 0.005**	**31.631 ± 0.004**	**39.118 ± 0.003**

22

Table 10: APSS results comparing `CCP`, `Cluster-CP`, and `RC3P` with ResNet-20 model under different imbalance ratio $\rho = 0.5$ and $\rho = 0.1$ where $\alpha = 0.01$. For a fair comparison of prediction set size, we set UCR of RC3P the same as or smaller (more restrictive) than that of `CCP` and `Cluster-CP` under 0.16 on CIFAR-10 and 0.03 on other datasets. The APSS results show that `RC3P` significantly outperforms `CCP` and `Cluster-CP` in terms of average prediction set size with 16.911% (four datasets) or 22.549% (excluding CIFAR-10) reduction in prediction size on average over min{`CCP`, `cluster-CP`}.

Conformity Score	Methods	EXP		POLY		MAJ	
		$\rho = 0.5$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.1$
		CIFAR-10					
APS	CCP	7.250 ± 0.164	$\mathbf{7.387 \pm 0.116}$	7.173 ± 0.079	7.596 ± 0.109	7.392 ± 0.128	$\mathbf{8.864 \pm 0.108}$
	Cluster-CP	$\mathbf{5.528 \pm 0.103}$	8.332 ± 0.060	6.954 ± 0.084	7.762 ± 0.143	7.586 ± 0.113	9.308 ± 0.054
	RC3P	5.671 ± 0.046	$\mathbf{7.387 \pm 0.116}$	$\mathbf{6.309 \pm 0.042}$	$\mathbf{7.276 \pm 0.010}$	$\mathbf{6.779 \pm 0.013}$	$\mathbf{8.864 \pm 0.108}$
RAPS	CCP	7.294 ± 0.160	$\mathbf{7.458 \pm 0.101}$	7.067 ± 0.106	7.597 ± 0.096	7.547 ± 0.134	$\mathbf{8.884 \pm 0.106}$
	Cluster-CP	$\mathbf{5.568 \pm 0.103}$	8.288 ± 0.118	6.867 ± 0.078	7.795 ± 0.136	7.813 ± 0.142	9.239 ± 0.055
	RC3P	5.673 ± 0.040	$\mathbf{7.458 \pm 0.101}$	$\mathbf{6.310 \pm 0.046}$	$\mathbf{7.253 \pm 0.006}$	$\mathbf{6.780 \pm 0.015}$	$\mathbf{8.884 \pm 0.106}$
		CIFAR-100					
APS	CCP	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$
	Cluster-CP	65.523 ± 0.495	69.063 ± 0.512	67.012 ± 0.739	81.997 ± 0.390	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$
	RC3P	$\mathbf{55.621 \pm 0.007}$	$\mathbf{63.039 \pm 0.007}$	$\mathbf{60.258 \pm 0.005}$	$\mathbf{74.927 \pm 0.007}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$
RAPS	CCP	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$
	Cluster-CP	65.584 ± 0.508	69.373 ± 0.466	66.313 ± 0.745	82.043 ± 0.439	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$
	RC3P	$\mathbf{55.632 \pm 0.008}$	$\mathbf{63.021 \pm 0.006}$	$\mathbf{60.205 \pm 0.006}$	$\mathbf{74.885 \pm 0.006}$	$\mathbf{100.0 \pm 0.0}$	$\mathbf{100.0 \pm 0.0}$
		mini-ImageNet					
APS	CCP	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	Cluster-CP	74.019 ± 0.699	71.300 ± 0.674	75.546 ± 0.683	70.996 ± 0.702	74.508 ± 0.531	72.803 ± 0.536
	RC3P	$\mathbf{55.321 \pm 0.003}$	$\mathbf{54.214 \pm 0.004}$	$\mathbf{56.018 \pm 0.006}$	$\mathbf{53.732 \pm 0.004}$	$\mathbf{54.483 \pm 0.007}$	$\mathbf{53.522 \pm 0.005}$
RAPS	CCP	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	Cluster-CP	73.893 ± 0.734	70.638 ± 0.657	75.546 ± 0.683	71.098 ± 0.706	74.675 ± 0.578	73.345 ± 0.474
	RC3P	$\mathbf{55.270 \pm 0.003}$	$\mathbf{54.184 \pm 0.003}$	$\mathbf{56.733 \pm 0.006}$	$\mathbf{53.736 \pm 0.004}$	$\mathbf{55.304 \pm 0.004}$	$\mathbf{53.532 \pm 0.005}$
		Food-101					
APS	CCP	101.0 ± 0.0	101.0 ± 0.0	101.0 ± 0.0	101.0 ± 0.0	101.0 ± 0.0	101.0 ± 0.0
	Cluster-CP	81.489 ± 0.957	87.092 ± 0.588	82.257 ± 0.514	86.539 ± 0.453	83.293 ± 0.583	88.603 ± 0.401
	RC3P	$\mathbf{67.443 \pm 0.004}$	$\mathbf{57.055 \pm 0.005}$	$\mathbf{57.722 \pm 0.006}$	$\mathbf{62.931 \pm 0.005}$	$\mathbf{68.267 \pm 0.005}$	$\mathbf{65.413 \pm 0.005}$
RAPS	CCP	101.0 ± 0.0	101.0 ± 0.0	101.0 ± 0.0	101.0 ± 0.0	101.0 ± 0.0	101.0 ± 0.0
	Cluster-CP	81.505 ± 0.955	87.103 ± 0.587	82.272 ± 0.513	86.517 ± 0.455	83.367 ± 0.635	88.604 ± 0.404
	RC3P	$\mathbf{67.444 \pm 0.004}$	$\mathbf{57.069 \pm 0.005}$	$\mathbf{57.722 \pm 0.006}$	$\mathbf{62.938 \pm 0.004}$	$\mathbf{68.266 \pm 0.005}$	$\mathbf{65.457 \pm 0.006}$

23

C.8 Comparison Experiments when models are trained in different epochs

With the same loss function, training criteria, evaluation metrics, and two scoring functions, we add the comparison experiments with CCP, and Cluster-CP on four datasets with different imbalanced types and imbalance ratio $\rho = 0.5$ and $\rho = 0.1$ and $\alpha = 0.1$ when models are trained with 50 epochs. The overall performance is summarized in Table 11. We highlight that we also select the g from the set $g \in \{0.25, 0.5, 0.75, 1.0\}$ to find the minimal g that CCP, Cluster-CP, and RC3P approximately achieves the target class conditional coverage.

Based on the results in Table 7, we make the following observations: (i) CCP, Cluster-CP, and RC3P can guarantee the class-conditional coverage; and (ii) RC3P significantly outperforms CCP and Cluster-CP on three datasets by producing smaller prediction sets.

Table 11: APSS results comparing CCP, Cluster-CP, and RC3P with ResNet-20 model under different imbalance ratio $\rho = 0.5$ and $\rho = 0.1$ where $\alpha = 0.1$ and models are trained with 50 epochs. For a fair comparison of prediction set size, we set UCR of RC3P the same as or smaller (more restrictive) than that of CCP and Cluster-CP under 0.16 on CIFAR-10 and 0.03 on other datasets. The APSS results show that RC3P significantly outperforms CCP and Cluster-CP in terms of average prediction set size with 21.441% (four datasets) or 28.588% (excluding CIFAR-10) reduction in prediction size on average over $\min\{\text{CCP}, \text{cluster-CP}\}$.

Conformity Score	Methods	EXP		POLY		MAJ	
		$\rho = 0.5$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.1$
			CIFAR-10				
APS	CCP	**2.420 ± 0.019**	**2.661 ± 0.015**	**2.399 ± 0.013**	**2.519 ± 0.022**	**2.651 ± 0.031**	**4.053 ± 0.021**
	Cluster-CP	4.006 ± 0.019	3.574 ± 0.023	3.144 ± 0.020	2.994 ± 0.029	3.698 ± 0.044	5.290 ± 0.016
	RC3P	**2.420 ± 0.019**	**2.661 ± 0.015**	**2.399 ± 0.013**	**2.519 ± 0.022**	**2.651 ± 0.031**	**4.053 ± 0.021**
RAPS	CCP	**2.096 ± 0.014**	**2.533 ± 0.019**	**2.383 ± 0.026**	**2.247 ± 0.017**	**2.232 ± 0.019**	**3.233 ± 0.021**
	Cluster-CP	2.625 ± 0.017	3.099 ± 0.021	2.840 ± 0.043	2.843 ± 0.026	2.770 ± 0.025	3.961 ± 0.029
	RC3P	**2.096 ± 0.014**	**2.533 ± 0.019**	**2.383 ± 0.026**	**2.247 ± 0.017**	**2.232 ± 0.019**	**3.233 ± 0.021**
			CIFAR-100				
APS	CCP	52.655 ± 0.473	52.832 ± 0.308	54.523 ± 0.441	61.768 ± 0.195	52.119 ± 0.197	58.333 ± 0.299
	Cluster-CP	42.990 ± 0.655	43.275 ± 0.833	44.114 ± 0.458	58.226 ± 0.627	39.841 ± 0.836	53.409 ± 0.520
	RC3P	**24.872 ± 0.008**	**25.107 ± 0.006**	**27.757 ± 0.004**	**35.733 ± 0.010**	**24.496 ± 0.010**	**32.172 ± 0.007**
RAPS	CCP	52.662 ± 0.473	52.841 ± 0.307	54.528 ± 0.442	61.766 ± 0.195	52.129 ± 0.197	58.331 ± 0.299
	Cluster-CP	43.024 ± 0.648	43.277 ± 0.839	44.120 ± 0.458	58.212 ± 0.629	39.864 ± 0.845	53.402 ± 0.518
	RC3P	**24.872 ± 0.008**	**25.107 ± 0.006**	**27.757 ± 0.004**	**35.733 ± 0.010**	**24.496 ± 0.010**	**32.173 ± 0.007**
			mini-ImageNet				
APS	CCP	42.404 ± 0.213	41.154 ± 0.191	38.433 ± 0.248	36.363 ± 0.228	36.047 ± 0.191	37.600 ± 0.208
	Cluster-CP	42.006 ± 0.430	41.101 ± 0.224	39.016 ± 0.273	36.046 ± 0.467	35.721 ± 0.355	37.975 ± 0.559
	RC3P	**32.022 ± 0.005**	**31.909 ± 0.004**	**28.460 ± 0.003**	**26.383 ± 0.003**	**26.128 ± 0.005**	**28.127 ± 0.005**
RAPS	CCP	42.516 ± 0.215	37.552 ± 0.192	38.730 ± 0.218	37.800 ± 0.186	36.595 ± 0.244	36.057 ± 0.206
	Cluster-CP	42.231 ± 0.386	37.448 ± 0.332	38.602 ± 0.327	37.939 ± 0.309	36.351 ± 0.308	35.724 ± 0.242
	RC3P	**32.022 ± 0.005**	**29.114 ± 0.004**	**28.197 ± 0.006**	**27.626 ± 0.004**	**25.853 ± 0.003**	**25.948 ± 0.003**
			Food-101				
APS	CCP	41.669 ± 0.118	51.395 ± 0.247	44.261 ± 0.165	58.816 ± 0.162	52.672 ± 0.169	57.312 ± 0.162
	Cluster-CP	44.883 ± 0.336	54.684 ± 0.475	47.794 ± 0.420	60.727 ± 0.178	56.100 ± 0.257	60.200 ± 0.543
	RC3P	**31.987 ± 0.005**	**36.118 ± 0.016**	**34.576 ± 0.006**	**49.299 ± 0.005**	**43.680 ± 0.005**	**47.649 ± 0.006**
RAPS	CCP	41.803 ± 0.157	48.548 ± 0.107	44.288 ± 0.165	56.592 ± 0.165	47.264 ± 0.120	56.666 ± 0.160
	Cluster-CP	44.810 ± 0.565	51.091 ± 0.375	47.861 ± 0.428	59.262 ± 0.306	50.211 ± 0.474	60.183 ± 0.507
	RC3P	**34.240 ± 0.115**	**36.425 ± 0.024**	**34.576 ± 0.006**	**46.074 ± 0.004**	**37.055 ± 0.006**	**48.012 ± 0.076**

C.9 Comparison Experiments with UCG metrics

We add the experiments without controlling coverage on imbalanced datasets under the same setting as the main paper. We then use the total under coverage gap (UCG, $\downarrow$ better) between class conditional coverage and target coverage $1 - \alpha$ of all under covered classes. We choose UCG as the fine-grained metric to differentiate the coverage performance in our experiment setting. Conditioned on similar APSS of all methods, RC3P significantly outperforms the best baselines with 35.18%(four datasets) or 46.91% (excluding CIFAR-10)reduction in UCG on average.

24

Table 12: UCG and APSS results comparing CCP, Cluster-CP, and RC3P with ResNet-20 model trained with 200 epochs under different imbalance types with imbalance ratio $\rho = 0.1$, where the coverage of each method are not aligned. The APSS results show that RC3P outperforms CCP and Cluster-CP in terms of average prediction set size with 1.64%(four datasets) or 2.19% (excluding CIFAR-10) reduction in prediction size on average over min{CCP, cluster-CP}. The UCG results show that RC3P achieve the similar class conditional coverage as CCP and Cluster-CP in terms of with 35.18%(four datasets) or 46.91% (excluding CIFAR-10) increment in the proportion of under coverage classes on average over min{CCP, cluster-CP}.

Conformity Score	Methods	EXP		POLY		MAJ	
		UCG	APSS	UCG	APSS	UCG	APSS
CIFAR-10							
APS	CCP	$\mathbf{0.014 \pm 0.000}$	1.573 ± 0.009	$\mathbf{0.032 \pm 0.000}$	1.494 ± 0.015	$\mathbf{0.068 \pm 0.000}$	2.175 ± 0.019
	Cluster-CP	0.166 ± 0.000	$\mathbf{1.438 \pm 0.012}$	0.124 ± 0.000	$\mathbf{1.280 \pm 0.007}$	0.144 ± 0.000	$\mathbf{2.079 \pm 0.023}$
	RC3P	$\mathbf{0.014 \pm 0.000}$	1.573 ± 0.009	$\mathbf{0.032 \pm 0.043}$	1.494 ± 0.015	$\mathbf{0.068 \pm 0.031}$	2.175 ± 0.019
RAPS	CCP	$\mathbf{0.014 \pm 0.000}$	1.573 ± 0.009	$\mathbf{0.032 \pm 0.000}$	1.494 ± 0.015	$\mathbf{0.070 \pm 0.000}$	2.179 ± 0.019
	Cluster-CP	0.166 ± 0.000	$\mathbf{1.438 \pm 0.012}$	0.124 ± 0.000	$\mathbf{1.280 \pm 0.007}$	0.144 ± 0.000	$\mathbf{2.079 \pm 0.023}$
	RC3P	$\mathbf{0.014 \pm 0.050}$	1.573 ± 0.009	$\mathbf{0.032 \pm 0.000}$	1.494 ± 0.015	$\mathbf{0.070 \pm 0.000}$	2.179 ± 0.019
CIFAR-100							
APS	CCP	1.920 ± 0.000	16.721 ± 0.174	2.000 ± 0.000	26.831 ± 0.150	2.400 ± 0.000	26.211 ± 0.216
	Cluster-CP	1.500 ± 0.000	15.657 ± 0.417	2.580 ± 0.000	26.709 ± 0.422	2.660 ± 0.000	25.145 ± 0.385
	RC3P	$\mathbf{0.840 \pm 0.000}$	$\mathbf{14.642 \pm 0.005}$	$\mathbf{1.200 \pm 0.000}$	$\mathbf{24.480 \pm 0.004}$	$\mathbf{1.460 \pm 0.000}$	$\mathbf{23.332 \pm 0.006}$
RAPS	CCP	1.920 ± 0.000	16.724 ± 0.174	2.020 ± 0.000	26.817 ± 0.150	2.400 ± 0.007	26.199 ± 0.216
	Cluster-CP	1.500 ± 0.000	15.767 ± 0.410	2.760 ± 0.000	26.712 ± 0.512	2.480 ± 0.000	25.153 ± 0.250
	RC3P	$\mathbf{0.840 \pm 0.000}$	$\mathbf{14.642 \pm 0.005}$	$\mathbf{1.200 \pm 0.000}$	$\mathbf{24.480 \pm 0.004}$	$\mathbf{1.460 \pm 0.000}$	$\mathbf{23.332 \pm 0.006}$
mini-ImageNet							
APS	CCP	1.486 ± 0.000	10.525 ± 0.093	1.620 ± 0.000	11.188 ± 0.094	1.280 ± 0.000	10.642 ± 0.055
	Cluster-CP	1.313 ± 0.000	11.133 ± 0.118	1.453 ± 0.000	11.547 ± 0.129	1.640 ± 0.000	11.186 ± 0.151
	RC3P	$\mathbf{0.713 \pm 0.000}$	$\mathbf{10.360 \pm 0.042}$	$\mathbf{0.653 \pm 0.000}$	$\mathbf{11.089 \pm 0.052}$	$\mathbf{0.600 \pm 0.000}$	$\mathbf{10.545 \pm 0.029}$
RAPS	CCP	1.526 ± 0.000	10.570 ± 0.093	1.620 ± 0.000	11.250 ± 0.095	1.293 ± 0.000	10.702 ± 0.055
	Cluster-CP	1.480 ± 0.000	11.192 ± 0.123	1.513 ± 0.000	11.704 ± 0.124	1.586 ± 0.000	11.231 ± 0.156
	RC3P	$\mathbf{0.713 \pm 0.000}$	$\mathbf{10.377 \pm 0.035}$	$\mathbf{0.653 \pm 0.000}$	$\mathbf{11.126 \pm 0.046}$	$\mathbf{0.600 \pm 0.000}$	$\mathbf{10.571 \pm 0.021}$
Food-101							
APS	CCP	1.176 ± 0.000	14.019 ± 0.064	1.208 ± 0.000	17.288 ± 0.075	1.748 ± 0.000	17.663 ± 0.076
	Cluster-CP	1.296 ± 0.000	13.998 ± 0.107	1.704 ± 0.000	17.300 ± 0.183	2.148 ± 0.000	17.410 ± 0.130
	RC3P	$\mathbf{0.556 \pm 0.000}$	$\mathbf{13.564 \pm 0.003}$	$\mathbf{0.664 \pm 0.000}$	$\mathbf{16.608 \pm 0.006}$	$\mathbf{0.924 \pm 0.000}$	$\mathbf{16.890 \pm 0.005}$
RAPS	CCP	1.160 ± 0.000	14.019 ± 0.064	1.208 ± 0.000	17.301 ± 0.075	1.764 ± 0.000	17.679 ± 0.076
	Cluster-CP	1.308 ± 0.000	14.080 ± 0.113	1.804 ± 0.000	17.370 ± 0.198	1.944 ± 0.000	17.488 ± 0.138
	RC3P	$\mathbf{0.556 \pm 0.000}$	$\mathbf{13.564 \pm 0.003}$	$\mathbf{0.664 \pm 0.000}$	$\mathbf{16.608 \pm 0.006}$	$\mathbf{0.924 \pm 0.000}$	$\mathbf{16.890 \pm 0.005}$

25

C.10 Complete Experiment Results on Imbalanced Datasets

In this subsection, we report complete experimental results over four imbalanced datasets, three decaying types, and five imbalance ratios when epoch $= 200$ and $\alpha = 0.1$. Specifically, Table 13, 14, 15 report results on CIFAR-10 with three decaying types. Table 16, 17, 18 report results on CIFAR-100 with three decaying types. Table 19, 20, 21 report results on mini-ImageNet with three decaying types. Table 22, 23, 24 report results on Food-101 with three decaying types.

Figure 5, Figure 6, Figure 7, Figure 8 and Figure 9 show the class-conditional coverage and the corresponding prediction set sizes on EXP $\rho = 0.5$, POLY $\rho = 0.1$, POLY $\rho = 0.5$, MAJ $\rho = 0.1$, MAJ $\rho = 0.5$, respectively. This result on EXP $\rho = 0.1$ is in Figure 1.

Figure 10, Figure 11, Figure 12, Figure 13 and Figure 14 illustrates the normalized frequency distribution of label ranks included in the prediction sets on EXP $\rho = 0.5$, POLY $\rho = 0.1$, POLY $\rho = 0.5$, MAJ $\rho = 0.1$, MAJ $\rho = 0.5$, respectively. This result on EXP $\rho = 0.1$ is in Figure 2. It is evident that the distribution of label ranks in the prediction set generated by `RC3P` tends to be lower compared to those produced by `CCP` and `Cluster-CP`. Furthermore, the probability density function tail for label ranks in the `RC3P` prediction set is notably shorter than that of other methods. This indicates that `RC3P` more effectively incorporates lower-ranked labels into prediction sets, as a result of its augmented rank calibration scheme.

Figure 15, Figure 16, Figure 17, Figure 18 and Figure 19 verify the condition numbers σ_y when models are fully trained (epoch $= 200$) on EXP $\rho = 0.5$, POLY $\rho = 0.1$, POLY $\rho = 0.5$, MAJ $\rho = 0.1$, MAJ $\rho = 0.5$, respectively. This result on EXP $\rho = 0.1$ is in Figure Figure 3. We also evaluate the condition numbers σ_y when models are lessly trained (epoch $= 50$) and $\alpha = 0.1$on EXP $\rho = 0.5$, EXP $\rho = 0.1$, POLY $\rho = 0.1$, POLY $\rho = 0.5$, MAJ $\rho = 0.1$, MAJ $\rho = 0.5$, respectively. These results are shown from Figure 21 to Figure 25. These results verify the validity of Lemma 4.2 and Equation 6 and confirm that the optimized trade-off between the coverage with inflated quantile and the constraint with calibrated rank leads to smaller prediction sets. They also show a stronger condition ($\sigma_y \leq 1$ for all y) than the weighted aggregation condition in (5). They also confirm that the condition number $\{\sigma_y\}_{y=1}^{C}$ could be evaluated on calibration datasets without testing datasets and thus decreases the computation cost. We notice that `RC3P` degenerates to CCP on CIFAR-10, so $\sigma_y = 1$ for all y and there is no trade-off. On the other three datasets, we observe significant conditions for the optimized trade-off in `RC3P`.

Table 13: Results comparing `CCP`, `cluster-CP`, and `RC3P` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type EXP and two scoring functions, APS and RAPS, on dataset CIFAR-10. We set UCR of `RC3P` the same as or better than that of `CCP` and `Cluster-CP` for a fair comparison of prediction set size.

Scoring function	Measure	Methods	EXP $\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
APS	UCR	CCP	0.050 ± 0.016	0.06 ± 0.021	0.050 ± 0.016	0.050 ± 0.021	0.100 ± 0.020
		Cluster-CP	$\mathbf{0.010 \pm 0.009}$	$\mathbf{0.050 \pm 0.021}$	$\mathbf{0.0 \pm 0.0}$	$\mathbf{0.030 \pm 0.015}$	$\mathbf{0.090 \pm 0.009}$
		RC3P	0.050 ± 0.016	0.06 ± 0.021	0.050 ± 0.016	0.050 ± 0.021	0.100 ± 0.020
	APSS	CCP	$\mathbf{1.555 \pm 0.010}$	$\mathbf{1.595 \pm 0.013}$	$\mathbf{1.643 \pm 0.008}$	$\mathbf{1.676 \pm 0.014}$	$\mathbf{1.855 \pm 0.014}$
		Cluster-CP	1.714 ± 0.018	1.745 ± 0.018	1.825 ± 0.014	1.901 ± 0.022	2.162 ± 0.015
		RC3P	$\mathbf{1.555 \pm 0.010}$	$\mathbf{1.595 \pm 0.013}$	$\mathbf{1.643 \pm 0.008}$	$\mathbf{1.676 \pm 0.014}$	$\mathbf{1.855 \pm 0.014}$
RAPS	UCR	CCP	0.050 ± 0.016	0.060 ± 0.021	0.050 ± 0.016	0.050 ± 0.021	$\mathbf{0.010 \pm 0.020}$
		Cluster-CP	$\mathbf{0.010 \pm 0.010}$	$\mathbf{0.050 \pm 0.021}$	$\mathbf{0.000 \pm 0.000}$	$\mathbf{0.030 \pm 0.014}$	$\mathbf{0.010 \pm 0.010}$
		RC3P	0.050 ± 0.016	0.060 ± 0.021	0.050 ± 0.016	0.050 ± 0.021	$\mathbf{0.010 \pm 0.020}$
	APSS	CCP	$\mathbf{1.555 \pm 0.010}$	$\mathbf{1.595 \pm 0.013}$	$\mathbf{1.643 \pm 0.008}$	$\mathbf{1.676 \pm 0.014}$	$\mathbf{1.855 \pm 0.014}$
		Cluster-CP	1.714 ± 0.018	1.745 ± 0.018	1.825 ± 0.014	1.901 ± 0.022	2.162 ± 0.015
		RC3P	$\mathbf{1.555 \pm 0.010}$	$\mathbf{1.595 \pm 0.013}$	$\mathbf{1.643 \pm 0.008}$	$\mathbf{1.676 \pm 0.014}$	$\mathbf{1.855 \pm 0.014}$

26

Table 14: Results comparing `CCP`, `cluster-CP`, and `RC3P` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type POLY and two scoring functions, APS and RAPS, on dataset CIFAR-10. We set UCR of `RC3P` the same as or better than that of `CCP` and `Cluster-CP` for a fair comparison of prediction set size.

Scoring function	Measure	Methods	POLY				
			$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
APS	UCR	CCP	0.100 ± 0.028	0.060 ± 0.026	0.060 ± 0.015	$\mathbf{0.050 \pm 0.021}$	$\mathbf{0.050 \pm 0.021}$
		Cluster-CP	$\mathbf{0.080 \pm 0.019}$	$\mathbf{0.050 \pm 0.021}$	$\mathbf{0.050 \pm 0.025}$	$\mathbf{0.050 \pm 0.016}$	0.060 ± 0.015
		RC3P	0.100 ± 0.028	0.060 ± 0.026	0.060 ± 0.015	$\mathbf{0.050 \pm 0.021}$	$\mathbf{0.050 \pm 0.021}$
	APSS	CCP	$\mathbf{1.538 \pm 0.010}$	$\mathbf{1.546 \pm 0.011}$	$\mathbf{1.580 \pm 0.014}$	$\mathbf{1.627 \pm 0.011}$	$\mathbf{1.776 \pm 0.012}$
		Cluster-CP	1.706 ± 0.014	1.718 ± 0.014	1.758 ± 0.016	1.783 ± 0.016	1.928 ± 0.013
		RC3P	$\mathbf{1.538 \pm 0.010}$	$\mathbf{1.546 \pm 0.011}$	$\mathbf{1.580 \pm 0.014}$	$\mathbf{1.627 \pm 0.011}$	$\mathbf{1.776 \pm 0.012}$
RAPS	UCR	CCP	0.100 ± 0.028	0.060 ± 0.025	0.060 ± 0.016	$\mathbf{0.050 \pm 0.021}$	$\mathbf{0.050 \pm 0.021}$
		Cluster-CP	$\mathbf{0.080 \pm 0.019}$	$\mathbf{0.050 \pm 0.021}$	$\mathbf{0.050 \pm 0.025}$	$\mathbf{0.050 \pm 0.016}$	0.060 ± 0.015
		RC3P	0.100 ± 0.028	0.060 ± 0.025	0.060 ± 0.016	$\mathbf{0.050 \pm 0.021}$	$\mathbf{0.050 \pm 0.021}$
	APSS	CCP	$\mathbf{1.538 \pm 0.010}$	$\mathbf{1.546 \pm 0.011}$	$\mathbf{1.581 \pm 0.014}$	$\mathbf{1.627 \pm 0.011}$	$\mathbf{1.776 \pm 0.012}$
		Cluster-CP	1.706 ± 0.014	1.719 ± 0.014	1.759 ± 0.016	1.783 ± 0.016	1.929 ± 0.013
		RC3P	$\mathbf{1.538 \pm 0.010}$	$\mathbf{1.546 \pm 0.011}$	$\mathbf{1.581 \pm 0.014}$	$\mathbf{1.627 \pm 0.011}$	$\mathbf{1.776 \pm 0.012}$

Table 15: Results comparing `CCP`, `cluster-CP`, and `RC3P` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type MAJ and two scoring functions, APS and RAPS, on dataset CIFAR-10. We set UCR of `RC3P` the same as or better than that of `CCP` and `Cluster-CP` for a fair comparison of prediction set size.

Scoring function	Measure	Methods	MAJ				
			$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
APS	UCR	CCP	0.070 ± 0.014	0.050 ± 0.016	0.080 ± 0.019	0.070 ± 0.025	$\mathbf{0.040 \pm 0.015}$
		Cluster-CP	$\mathbf{0.020 \pm 0.012}$	$\mathbf{0.040 \pm 0.015}$	$\mathbf{0.020 \pm 0.013}$	$\mathbf{0.010 \pm 0.010}$	0.070 ± 0.014
		RC3P	0.070 ± 0.014	0.050 ± 0.016	0.080 ± 0.019	0.070 ± 0.025	$\mathbf{0.040 \pm 0.015}$
	APSS	CCP	$\mathbf{1.84 \pm 0.020}$	$\mathbf{1.825 \pm 0.014}$	$\mathbf{1.939 \pm 0.016}$	$\mathbf{2.054 \pm 0.013}$	$\mathbf{2.629 \pm 0.013}$
		Cluster-CP	1.948 ± 0.023	1.999 ± 0.027	2.167 ± 0.030	2.457 ± 0.021	3.220 ± 0.020
		RC3P	$\mathbf{1.84 \pm 0.020}$	$\mathbf{1.825 \pm 0.014}$	$\mathbf{1.939 \pm 0.016}$	$\mathbf{2.054 \pm 0.013}$	$\mathbf{2.629 \pm 0.013}$
RAPS	UCR	CCP	0.070 ± 0.014	0.050 ± 0.016	0.080 ± 0.019	0.070 ± 0.025	$\mathbf{0.040 \pm 0.015}$
		Cluster-CP	$\mathbf{0.020 \pm 0.013}$	$\mathbf{0.040 \pm 0.015}$	$\mathbf{0.020 \pm 0.012}$	$\mathbf{0.010 \pm 0.010}$	0.070 ± 0.014
		RC3P	0.070 ± 0.014	0.050 ± 0.016	0.080 ± 0.019	0.070 ± 0.025	$\mathbf{0.040 \pm 0.015}$
	APSS	CCP	$\mathbf{1.840 \pm 0.020}$	$\mathbf{1.825 \pm 0.014}$	$\mathbf{1.940 \pm 0.016}$	$\mathbf{2.055 \pm 0.013}$	$\mathbf{2.632 \pm 0.012}$
		Cluster-CP	1.948 ± 0.023	1.999 ± 0.028	2.168 ± 0.030	2.458 ± 0.021	3.219 ± 0.030
		RC3P	$\mathbf{1.840 \pm 0.020}$	$\mathbf{1.825 \pm 0.014}$	$\mathbf{1.940 \pm 0.016}$	$\mathbf{2.055 \pm 0.013}$	$\mathbf{2.632 \pm 0.012}$

Table 16: Results comparing `CCP`, `cluster-CP`, and `RC3P` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type EXP and two scoring functions, APS and RAPS, on dataset CIFAR-100. We set UCR of `RC3P` the same as or better than that of `CCP` and `Cluster-CP` for a fair comparison of prediction set size.

Scoring function	Measure	Methods	EXP				
			$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
APS	UCR	CCP	0.007 ± 0.002	0.017 ± 0.004	0.012 ± 0.004	0.015 ± 0.003	$\mathbf{0.010 \pm 0.002}$
		Cluster-CP	0.012 ± 0.002	0.012 ± 0.003	$\mathbf{0.006 \pm 0.002}$	0.035 ± 0.008	0.016 ± 0.004
		RC3P	$\mathbf{0.005 \pm 0.002}$	$\mathbf{0.009 \pm 0.001}$	0.011 ± 0.003	$\mathbf{0.013 \pm 0.003}$	0.011 ± 0.002
	APSS	CCP	44.224 ± 0.341	44.486 ± 0.420	47.672 ± 0.463	46.955 ± 0.402	50.969 ± 0.345
		Cluster-CP	29.238 ± 0.609	30.602 ± 0.553	32.126 ± 0.563	33.714 ± 0.863	37.592 ± 0.857
		RC3P	$\mathbf{17.705 \pm 0.004}$	$\mathbf{18.311 \pm 0.005}$	$\mathbf{19.608 \pm 0.007}$	$\mathbf{20.675 \pm 0.005}$	$\mathbf{21.954 \pm 0.005}$
RAPS	UCR	CCP	0.007 ± 0.002	0.017 ± 0.004	0.012 ± 0.003	0.015 ± 0.003	$\mathbf{0.011 \pm 0.002}$
		Cluster-CP	0.011 ± 0.003	$\mathbf{0.009 \pm 0.002}$	$\mathbf{0.006 \pm 0.002}$	0.034 ± 0.007	0.017 ± 0.004
		RC3P	$\mathbf{0.005 \pm 0.002}$	0.012 ± 0.003	0.011 ± 0.003	$\mathbf{0.013 \pm 0.003}$	$\mathbf{0.011 \pm 0.002}$
	APSS	CCP	44.250 ± 0.342	44.499 ± 0.420	47.688 ± 0.569	46.960 ± 0.404	50.970 ± 0.345
		Cluster-CP	29.267 ± 0.612	30.595 ± 0.549	32.161 ± 0.564	33.713 ± 0.864	37.595 ± 0.862
		RC3P	$\mathbf{17.705 \pm 0.004}$	$\mathbf{18.311 \pm 0.005}$	$\mathbf{19.609 \pm 0.007}$	$\mathbf{20.675 \pm 0.005}$	$\mathbf{21.954 \pm 0.005}$

27

Table 17: Results comparing `CCP`, `cluster-CP`, and `RC3P` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type POLY and two scoring functions, APS and RAPS, on dataset CIFAR-100. We set UCR of `RC3P` the same as or better than that of `CCP` and `Cluster-CP` for a fair comparison of prediction set size.

Scoring function	Measure	Methods	POLY				
			$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
APS	UCR	CCP	0.010± 0.002	0.008± 0.002	0.016± 0.003	0.012± 0.004	**0.014± 0.003**
		Cluster-CP	0.020± 0.003	0.020± 0.002	0.026± 0.004	**0.009± 0.003**	0.034± 0.005
		RC3P	**0.009± 0.003**	**0.005± 0.002**	**0.013± 0.004**	**0.011± 0.004**	0.015± 0.003
	APSS	CCP	49.889± 0.353	54.011± 0.466	56.031± 0.406	59.888± 0.255	64.343± 0.237
		Cluster-CP	38.252± 0.316	39.585± 0.545	43.310± 0.824	47.461± 0.979	52.391± 0.595
		RC3P	**23.048± 0.008**	**24.335± 0.005**	**26.366± 0.010**	**28.887± 0.006**	**33.829± 0.005**
RAPS	UCR	CCP	0.010± 0.002	0.008± 0.002	0.016± 0.003	0.012± 0.004	**0.015± 0.003**
		Cluster-CP	0.019± 0.004	0.020± 0.002	0.026± 0.005	**0.009± 0.003**	0.034± 0.005
		RC3P	**0.009± 0.003**	**0.005± 0.002**	**0.013± 0.004**	0.011± 0.004	**0.015± 0.003**
	APSS	CCP	49.886± 0.353	53.994± 0.467	56.020± 0.406	59.870± 0.253	64.332± 0.236
		Cluster-CP	38.258± 0.320	39.566± 0.549	43.304± 0.549	47.450± 0.969	52.374± 0.592
		RC3P	**23.048± 0.008**	**24.335± 0.005**	**26.366± 0.010**	**28.886± 0.006**	**33.185± 0.005**

Table 18: Results comparing `CCP`, `cluster-CP`, and `RC3P` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type MAJ and two scoring functions, APS and RAPS, on dataset CIFAR-100. We set UCR of `RC3P` the same as or better than that of `CCP` and `Cluster-CP` for a fair comparison of prediction set size.

Scoring function	Measure	Methods	MAJ				
			$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
APS	UCR	CCP	0.016± 0.003	**0.007± 0.002**	0.017± 0.004	**0.010± 0.002**	**0.008± 0.004**
		Cluster-CP	**0.008± 0.002**	0.012± 0.003	0.021± 0.004	0.021± 0.005	0.019± 0.005
		RC3P	0.016± 0.003	0.010± 0.003	**0.015± 0.004**	**0.010± 0.002**	**0.008± 0.004**
	APSS	CCP	44.194± 0.514	49.231± 0.129	53.676± 0.372	55.024± 0.254	64.642± 0.535
		Cluster-CP	31.518± 0.335	35.355± 0.563	37.514± 0.538	43.619± 0.600	50.883± 0.673
		RC3P	**18.581± 0.007**	**21.080± 0.010**	**22.606± 0.007**	**26.785± 0.007**	**32.699± 0.005**
RAPS	UCR	CCP	0.015± 0.003	**0.007± 0.002**	0.011± 0.004	**0.010± 0.003**	**0.008± 0.004**
		Cluster-CP	**0.008± 0.003**	0.011± 0.003	0.021± 0.004	0.021± 0.002	0.018± 0.005
		RC3P	0.015± 0.003	0.010± 0.003	**0.015± 0.004**	**0.010± 0.002**	**0.008± 0.004**
	APSS	CCP	48.343± 0.353	49.252± 0.128	53.666± 0.371	55.016± 0.254	64.633± 0.535
		Cluster-CP	31.513± 0.325	35.352± 0.547	37.503± 0.535	43.615± 0.608	50.379± 0.684
		RC3P	**18.581± 0.006**	**21.080± 0.010**	**22.605± 0.007**	**26.786± 0.007**	**32.699± 0.006**

Table 19: Results comparing `CCP`, `cluster-CP`, and `RC3P` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type EXP and two scoring function, APS and RAPS, on dataset mini-ImageNet. We set UCR of `RC3P` the same as or better than that of `CCP` and `Cluster-CP` for a fair comparison of prediction set size.

Scoring function	Measure	Methods	EXP				
			$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
APS	UCR	CCP	0.008± 0.004	0.003± 0.002	0.003± 0.001	0.003± 0.003	0.008± 0.004
		Cluster-CP	0.014± 0.004	0.005± 0.002	0.010± 0.002	0.010± 0.003	0.012± 0.004
		RC3P	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**	**0.001± 0.001**
	APSS	CCP	26.676± 0.171	25.663± 0.182	25.941± 0.180	26.127± 0.187	26.111± 0.194
		Cluster-CP	25.889± 0.301	25.878± 0.258	25.680± 0.294	25.522± 0.311	25.253± 0.346
		RC3P	**18.129± 0.003**	**17.546± 0.002**	**17.352± 0.003**	**17.006± 0.003**	**17.082± 0.002**
RAPS	UCR	CCP	0.008± 0.004	0.004± 0.003	0.003± 0.001	0.003± 0.003	0.009± 0.004
		Cluster-CP	0.006± 0.002	0.003± 0.001	0.009± 0.002	0.008± 0.003	0.013± 0.005
		RC3P	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**	**0.001± 0.001**
	APSS	CCP	26.756± 0.178	26.621± 0.182	25.021± 0.182	26.216± 0.188	26.212± 0.199
		Cluster-CP	26.027± 0.325	26.000± 0.283	25.922± 0.253	25.564± 0.358	25.415± 0.289
		RC3P	**18.129± 0.003**	**17.546± 0.002**	**17.352± 0.003**	**17.006± 0.003**	**17.082± 0.002**

28

Table 20: Results comparing `CCP`, `cluster-CP`, and `RC3P` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type POLY and two scoring function, APS and RAPS, on dataset mini-ImageNet. We set UCR of `RC3P` the same as or better than that of `CCP` and `Cluster-CP` for a fair comparison of prediction set size.

Scoring function	Measure	Methods	POLY				
			$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
APS	UCR	CCP	0.005± 0.002	0.004± 0.002	0.005± 0.002	0.002± 0.001	0.004± 0.001
		Cluster-CP	0.011± 0.003	0.013± 0.003	0.015± 0.004	0.012± 0.003	0.014± 0.003
		RC3P	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**
	APSS	CCP	26.626± 0.133	26.343± 0.214	27.168± 0.203	27.363± 0.252	26.159± 0.208
		Cluster-CP	26.150± 0.393	25.348± 0.231	26.132± 0.415	26.390± 0.270	25.633± 0.268
		RC3P	**17.784± 0.003**	**17.752± 0.003**	**17.652± 0.003**	**17.629± 0.003**	**17.465± 0.003**
RAPS	UCR	CCP	0.005± 0.002	0.004± 0.002	0.005± 0.002	0.002± 0.001	0.004± 0.002
		Cluster-CP	0.009± 0.003	0.016± 0.004	0.017± 0.004	0.009± 0.003	0.016± 0.003
		RC3P	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**
	APSS	CCP	26.689± 0.142	26.437± 0.213	27.254± 0.201	27.450± 0.249	26.248± 0.219
		Cluster-CP	26.288± 0.407	25.627± 0.318	26.220± 0.432	26.559± 0.242	25.712± 0.315
		RC3P	**17.784± 0.003**	**17.752± 0.003**	**17.652± 0.003**	**17.629± 0.003**	**17.465± 0.003**

Table 21: Results comparing `CCP`, `cluster-CP`, and `RC3P` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type MAJ and two scoring function, APS and RAPS, on dataset mini-ImageNet. We set UCR of `RC3P` the same as or better than that of `CCP` and `Cluster-CP` for a fair comparison of prediction set size.

Scoring function	Measure	Methods	MAJ				
			$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
APS	UCR	CCP	0.010± 0.004	0.009± 0.003	**0.0± 0.0**	0.005± 0.002	0.005± 0.002
		Cluster-CP	0.008± 0.002	0.010± 0.000	0.010± 0.003	0.012± 0.004	0.010± 0.003
		RC3P	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**
	APSS	CCP	27.313± 0.154	27.233± 0.246	26.939± 0.177	26.676± 0.267	25.629± 0.207
		Cluster-CP	26.918± 0.241	26.156± 0.255	25.786± 0.356	25.632± 0.383	25.348± 0.334
		RC3P	**18.111± 0.002**	**17.874± 0.002**	**18.081± 0.003**	**17.800± 0.002**	**17.167± 0.004**
RAPS	UCR	CCP	0.009± 0.003	0.009± 0.003	**0.0± 0.0**	0.005± 0.002	0.005± 0.002
		Cluster-CP	0.007± 0.002	0.011± 0.002	0.013± 0.004	0.014± 0.004	0.009± 0.002
		RC3P	**0.0±(0.0)**	**0.0±(0.0)**	**0.0±(0.0)**	**0.0±(0.0)**	**0.0±(0.0)**
	APSS	CCP	27.397± 0.162	27.320± 0.244	27.013± 0.177	26.782± 0.269	25.725± 0.214
		Cluster-CP	26.969± 0.305	26.293± 0.245	25.956± 0.308	25.803± 0.440	25.532± 0.350
		RC3P	**18.111± 0.002**	**17.874± 0.002**	**18.081± 0.003**	**17.800± 0.002**	**17.167± 0.004**

Table 22: Results comparing `CCP`, `cluster-CP`, and `RC3P` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type EXP and two scoring function, APS and RAPS, on dataset Food-101. We set UCR of `RC3P` the same as or better than that of `CCP` and `Cluster-CP` for a fair comparison of prediction set size.

Scoring function	Measure	Methods	EXP				
			$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
APS	UCR	CCP	0.006± 0.002	0.010± 0.002	0.008± 0.002	0.014± 0.004	0.006± 0.002
		Cluster-CP	0.003± 0.002	0.009± 0.003	0.006± 0.003	0.008± 0.003	0.009± 0.003
		RC3P	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**
	APSS	CCP	27.003± 0.183	27.024± 0.162	28.074± 0.199	28.512± 0.154	30.875± 0.163
		Cluster-CP	29.020± 0.281	30.120± 0.440	30.529± 0.381	31.096± 0.350	33.327± 0.440
		RC3P	**18.369± 0.003**	**18.339± 0.004**	**18.803± 0.003**	**19.612± 0.005**	**21.556± 0.006**
RAPS	UCR	CCP	0.006± 0.003	0.010± 0.002	0.008± 0.002	0.014± 0.004	0.006± 0.002
		Cluster-CP	0.004± 0.003	0.010± 0.003	0.006± 0.003	0.010± 0.002	0.012± 0.004
		RC3P	**0.0±(0.0)**	**0.0±(0.0)**	**0.0±(0.0)**	**0.0±(0.0)**	**0.0±(0.0)**
	APSS	CCP	27.022± 0.192	27.043± 0.163	28.098± 0.199	28.535± 0.155	30.900± 0.170
		Cluster-CP	28.953± 0.333	30.242± 0.466	30.587± 0.377	30.924± 0.317	33.375± 0.377
		RC3P	**18.369± 0.004**	**18.339± 0.004**	**18.803± 0.003**	**19.612± 0.005**	**21.556± 0.006**

29

Table 23: Results comparing `CCP`, `cluster-CP`, and `RC3P` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type POLY and two scoring function, APS and RAPS, on dataset Food-101. We set UCR of `RC3P` the same as or better than that of `CCP` and `Cluster-CP` for a fair comparison of prediction set size.

| Scoring function | Measure | Methods | POLY | | | | |
			$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
APS	UCR	CCP	0.009± 0.003	0.005± 0.003	0.009± 0.003	0.011± 0.003	0.008± 0.001
		Cluster-CP	0.004± 0.001	0.012± 0.002	0.012± 0.004	0.011± 0.002	0.009± 0.002
		RC3P	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**	**0.001± 0.001**
	APSS	CCP	30.943± 0.119	31.239± 0.198	32.283± 0.169	33.570± 0.163	35.912± 0.105
		Cluster-CP	33.079± 0.393	33.951± 0.531	34.626± 0.352	36.546± 0.490	38.301± 0.232
		RC3P	**21.499± 0.003**	**21.460± 0.005**	**22.882± 0.005**	**23.708± 0.004**	**25.853± 0.004**
RAPS	UCR	CCP	0.009± 0.003	0.006± 0.003	0.009± 0.003	0.011± 0.003	0.008± 0.001
		Cluster-CP	0.006± 0.002	0.013± 0.002	0.012± 0.004	0.016± 0.002	0.006± 0.003
		RC3P	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**	**0.001± 0.001**
	APSS	CCP	30.966± 0.125	31.257± 0.197	32.302± 0.169	33.595± 0.164	35.940± 0.111
		Cluster-CP	33.337± 0.409	33.936± 0.448	34.878± 0.282	36.505± 0.520	38.499± 0.216
		RC3P	**21.499± 0.003**	**21.460± 0.005**	**22.882± 0.005**	**23.708± 0.004**	**25.853± 0.004**

Table 24: Results comparing `CCP`, `cluster-CP`, and `RC3P` with ResNet-20 model under different imbalance ratio $\rho = 0.5$, $\rho = 0.4$, $\rho = 0.2$, and $\rho = 0.1$ with imbalance type MAJ and two scoring function, APS and RAPS, on dataset Food-101. We set UCR of `RC3P` the same as or better than that of `CCP` and `Cluster-CP` for a fair comparison of prediction set size.

| Scoring function | Measure | Methods | MAJ | | | | |
			$\rho = 0.5$	$\rho = 0.4$	$\rho = 0.3$	$\rho = 0.2$	$\rho = 0.1$
APS	UCR	CCP	0.006± 0.001	0.005± 0.002	0.008± 0.003	0.010± 0.002	0.008± 0.002
		Cluster-CP	0.011± 0.003	0.005± 0.002	0.014± 0.004	0.016± 0.004	0.011± 0.002
		RC3P	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**	**0.0± 0.0**
	APSS	CCP	27.415± 0.194	29.369± 0.120	30.672± 0.182	31.966± 0.165	36.776± 0.132
		Cluster-CP	30.071± 0.412	31.656± 0.261	32.857± 0.469	33.774± 0.494	39.632± 0.342
		RC3P	**19.398± 0.006**	**20.046± 0.004**	**21.425± 0.003**	**22.175± 0.004**	**26.585± 0.004**
RAPS	UCR	CCP	0.006± 0.002	0.005± 0.002	0.008± 0.003	0.010± 0.002	0.008± 0.002
		Cluster-CP	0.011± 0.003	0.005± 0.002	0.013± 0.004	0.014± 0.004	0.014± 0.004
		RC3P	**0.0±(0.0)**	**0.0±(0.0)**	**0.0±(0.0)**	**0.0±(0.0)**	**0.0±(0.0)**
	APSS	CCP	27.439± 0.203	29.393± 0.120	30.691± 0.182	31.987± 0.165	36.802± 0.138
		Cluster-CP	29.946± 0.407	31.409± 0.303	32.724± 0.551	33.686± 0.501	39.529± 0.306
		RC3P	**19.397± 0.006**	**20.046± 0.004**	**21.425± 0.003**	**22.175± 0.004**	**26.585± 0.004**

30
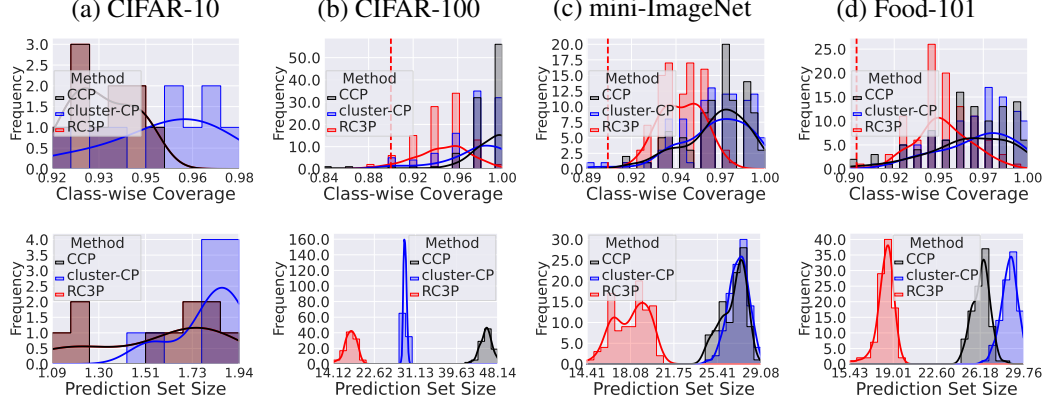
Figure 5: Class-conditional coverage (Top row) and prediction set size (Bottom row) achieved by CCP, Cluster-CP, and RC3P methods when $\alpha = 0.1$ on CIFAR-10, CIFAR-100, mini-ImageNet, and Food-101 datasets with imbalance type EXP for imbalance ratio $\rho = 0.5$. We clarify that RC3P overlaps with CCP on CIFAR-10. It is clear that RC3P has more densely distributed class-conditional coverage above 0.9 (the target $1 - \alpha$ class-conditional coverage) than CCP and Cluster-CP with significantly smaller prediction sets on CIFAR-100, mini-ImageNet and Food-101.
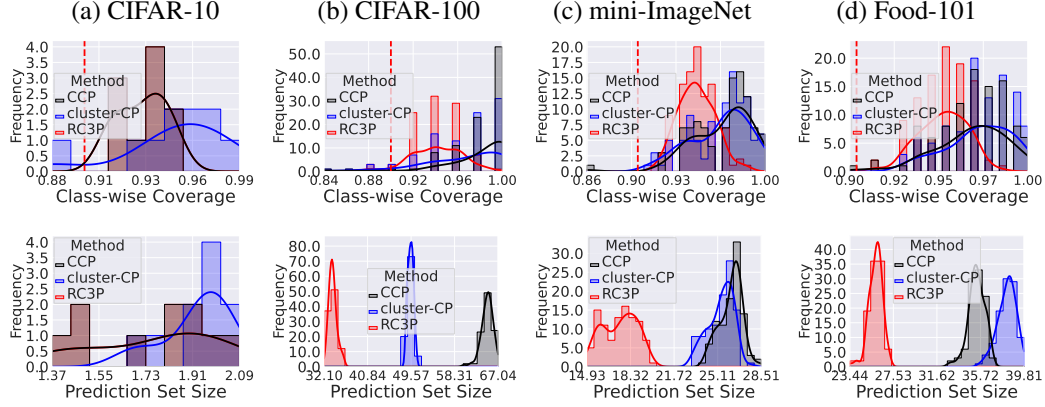
Figure 6: Class-conditional coverage (Top row) and prediction set size (Bottom row) achieved by CCP, Cluster-CP, and RC3P methods when $\alpha = 0.1$ on CIFAR-10, CIFAR-100, mini-ImageNet, and Food-101 datasets with imbalance type POLY for imbalance ratio $\rho = 0.1$. We clarify that RC3P overlaps with CCP on CIFAR-10. It is clear that RC3P has more densely distributed class-conditional coverage above 0.9 (the target $1 - \alpha$ class-conditional coverage) than CCP and Cluster-CP with significantly smaller prediction sets on CIFAR-100, mini-ImageNet and Food-101.

31
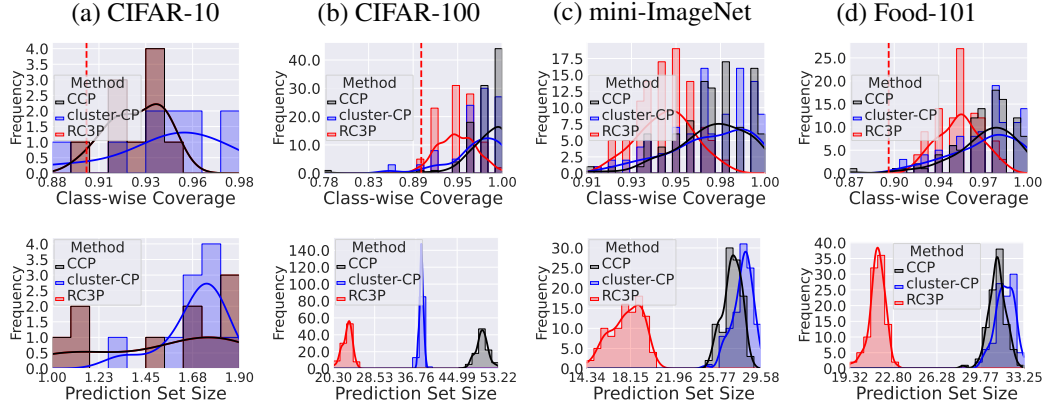
Figure 7: Class-conditional coverage (Top row) and prediction set size (Bottom row) achieved by CCP, Cluster-CP, and RC3P methods when $\alpha = 0.1$ on CIFAR-10, CIFAR-100, mini-ImageNet, and Food-101 datasets with imbalance type POLY for imbalance ratio $\rho = 0.5$. We clarify that RC3P overlaps with CCP on CIFAR-10. It is clear that RC3P has more densely distributed class-conditional coverage above 0.9 (the target $1 - \alpha$ class-conditional coverage) than CCP and Cluster-CP with significantly smaller prediction sets on CIFAR-100, mini-ImageNet and Food-101.
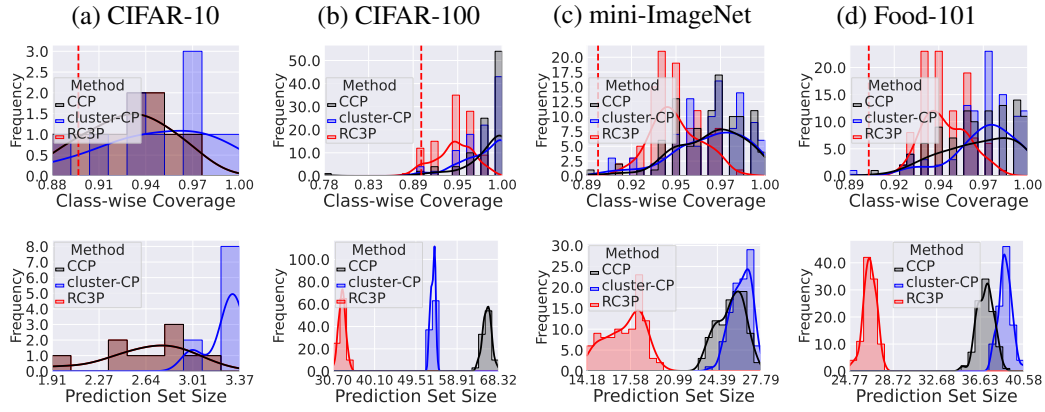
Figure 8: Class-conditional coverage (Top row) and prediction set size (Bottom row) achieved by CCP, Cluster-CP, and RC3P methods when $\alpha = 0.1$ on CIFAR-10, CIFAR-100, mini-ImageNet, and Food-101 datasets with imbalance type MAJ for imbalance ratio $\rho = 0.1$. We clarify that RC3P overlaps with CCP on CIFAR-10. It is clear that RC3P has more densely distributed class-conditional coverage above 0.9 (the target $1 - \alpha$ class-conditional coverage) than CCP and Cluster-CP with significantly smaller prediction sets on CIFAR-100, mini-ImageNet and Food-101.

32
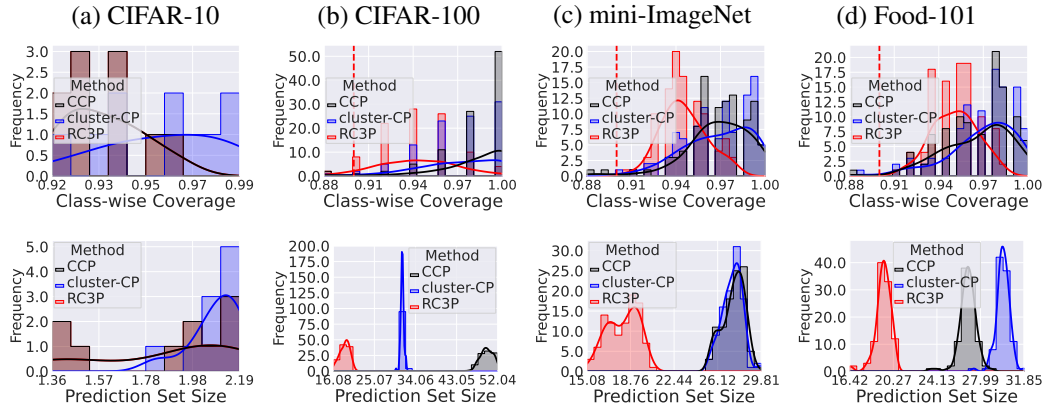
Figure 9: Class-conditional coverage (Top row) and prediction set size (Bottom row) achieved by CCP, Cluster-CP, and RC3P methods when $\alpha = 0.1$ on CIFAR-10, CIFAR-100, mini-ImageNet, and Food-101 datasets with imbalance type MAJ for imbalance ratio $\rho = 0.5$. We clarify that RC3P overlaps with CCP on CIFAR-10. It is clear that RC3P has more densely distributed class-conditional coverage above 0.9 (the target $1 - \alpha$ class-conditional coverage) than CCP and Cluster-CP with significantly smaller prediction sets on CIFAR-100, mini-ImageNet and Food-101.
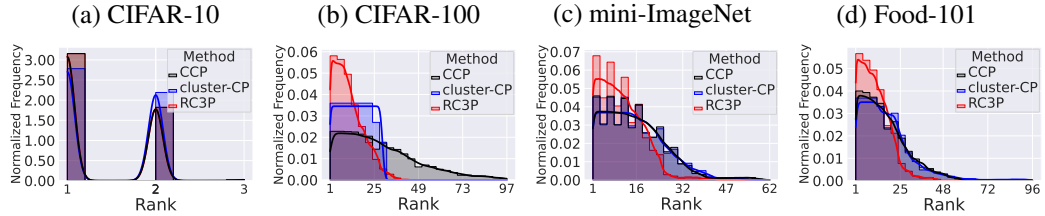
Figure 10: Visualization for the normalized frequency distribution of label ranks included in the prediction set of CCP, Cluster-CP, and RC3P with $\rho = 0.5$ EXP when $\alpha = 0.1$. It is clear that the distribution of normalized frequency generated by RC3P tends to be lower compared to those produced by CCP and Cluster-CP. Furthermore, the probability density function tail for label ranks in the RC3P prediction set is notably shorter than that of other methods.
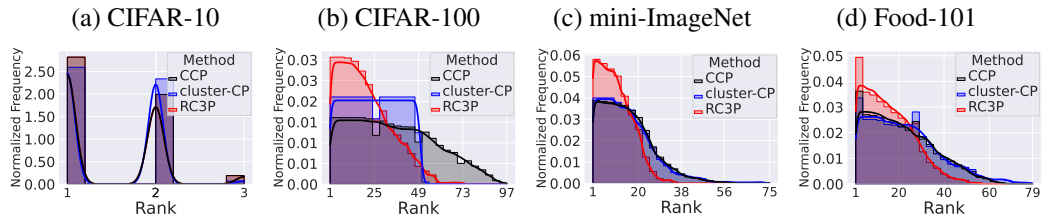
Figure 11: Visualization for the normalized frequency distribution of label ranks included in the prediction set of CCP, Cluster-CP, and RC3P with $\rho = 0.1$ POLY when $\alpha = 0.1$. It is clear that the distribution of normalized frequency generated by RC3P tends to be lower compared to those produced by CCP and Cluster-CP. Furthermore, the probability density function tail for label ranks in the RC3P prediction set is notably shorter than that of other methods.

33
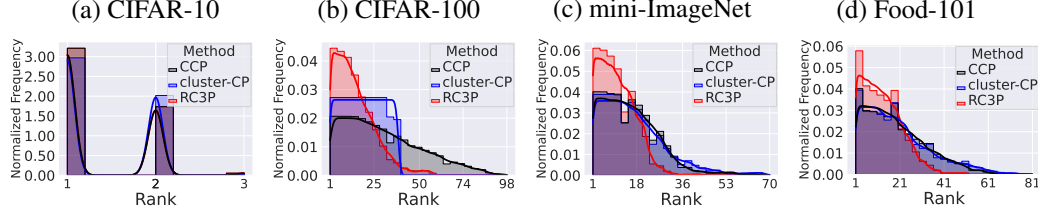
Figure 12: Visualization for the normalized frequency distribution of label ranks included in the prediction set of CCP, Cluster-CP, and RC3P with $\rho = 0.5$ POLY when $\alpha = 0.1$. It is clear that the distribution of normalized frequency generated by RC3P tends to be lower compared to those produced by CCP and Cluster-CP. Furthermore, the probability density function tail for label ranks in the RC3P prediction set is notably shorter than that of other methods.
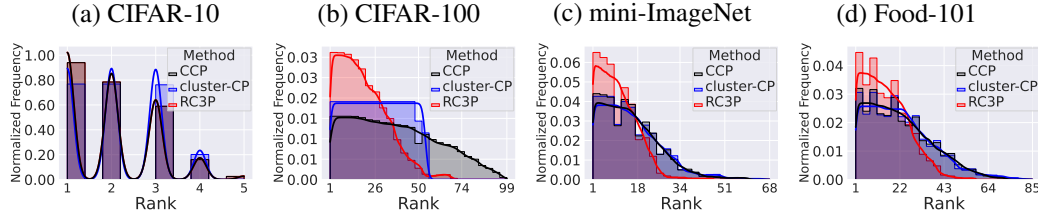
Figure 13: Visualization for the normalized frequency distribution of label ranks included in the prediction set of CCP, Cluster-CP, and RC3P with $\rho = 0.1$ MAJ when $\alpha = 0.1$. It is clear that the distribution of normalized frequency generated by RC3P tends to be lower compared to those produced by CCP and Cluster-CP. Furthermore, the probability density function tail for label ranks in the RC3P prediction set is notably shorter than that of other methods.
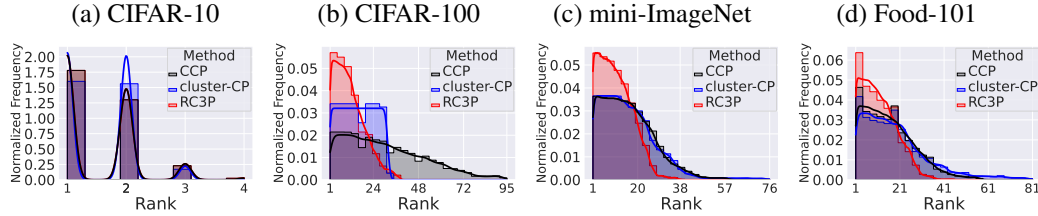
Figure 14: Visualization for the normalized frequency distribution of label ranks included in the prediction set of CCP, Cluster-CP, and RC3P with $\rho = 0.5$ MAJ when $\alpha = 0.1$. It is clear that the distribution of normalized frequency generated by RC3P tends to be lower compared to those produced by CCP and Cluster-CP. Furthermore, the probability density function tail for label ranks in the RC3P prediction set is notably shorter than that of other methods.
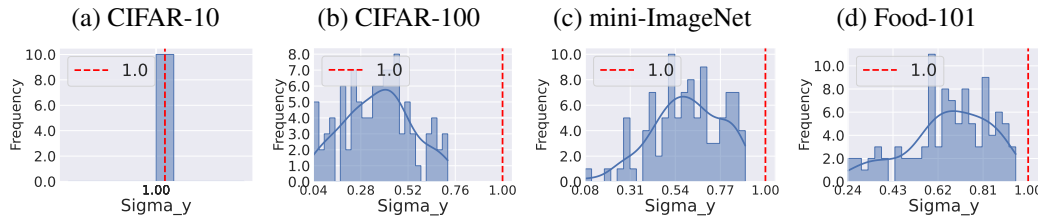
Figure 15: Verification of condition numbers $\{\sigma_y\}_{y=1}^{C}$ of Equation 6 when epoch $= 200$ and $\alpha = 0.1$ with $\rho = 0.5$ EXP. Vertical dashed lines represent the value 1, and we observe that all the condition numbers are smaller than 1. This verifies the validity of the condition for Lemma 4.2, and thus confirms that RC3P produces smaller prediction sets than CCP by the optimized trade-off between calibration on non-conformity scores and calibrated label ranks.

34

(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101
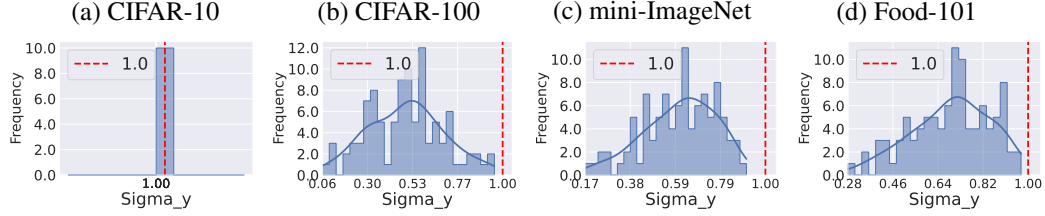
Figure 16: Verification of condition numbers $\{\sigma_y\}_{y=1}^C$ of Equation 6 when epoch $= 200$ and $\alpha = 0.1$ with $\rho = 0.1$ POLY. Vertical dashed lines represent the value 1, and we observe that all the condition numbers are smaller than 1. This verifies the validity of the condition for Lemma 4.2, and thus confirms that RC3P produces smaller prediction sets than CCP by the optimized trade-off between calibration on non-conformity scores and calibrated label ranks.

(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101
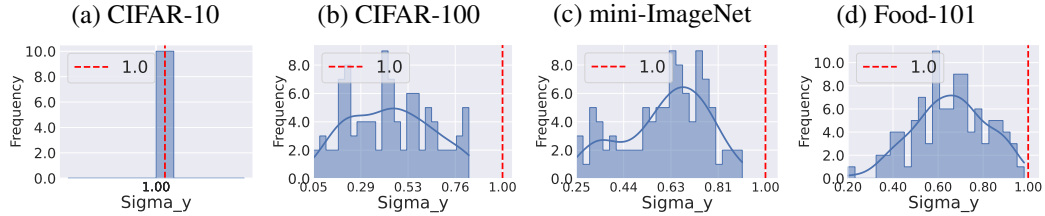
Figure 17: Verification of condition numbers $\{\sigma_y\}_{y=1}^C$ of Equation 6 when epoch $= 200$ and $\alpha = 0.1$ with $\rho = 0.5$ POLY. Vertical dashed lines represent the value 1, and we observe that all the condition numbers are smaller than 1. This verifies the validity of the condition for Lemma 4.2, and thus confirms that RC3P produces smaller prediction sets than CCP by the optimized trade-off between calibration on non-conformity scores and calibrated label ranks.

(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101
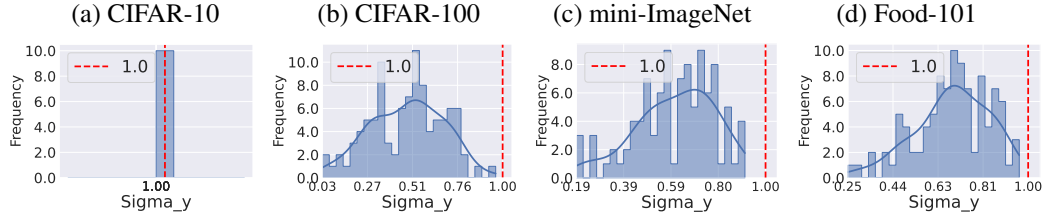
Figure 18: Verification of condition numbers $\{\sigma_y\}_{y=1}^C$ of Equation 6 when epoch $= 200$ and $\alpha = 0.1$ with $\rho = 0.1$ MAJ. Vertical dashed lines represent the value 1, and we observe that all the condition numbers are smaller than 1. This verifies the validity of the condition for Lemma 4.2, and thus confirms that RC3P produces smaller prediction sets than CCP by the optimized trade-off between calibration on non-conformity scores and calibrated label ranks.

(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101
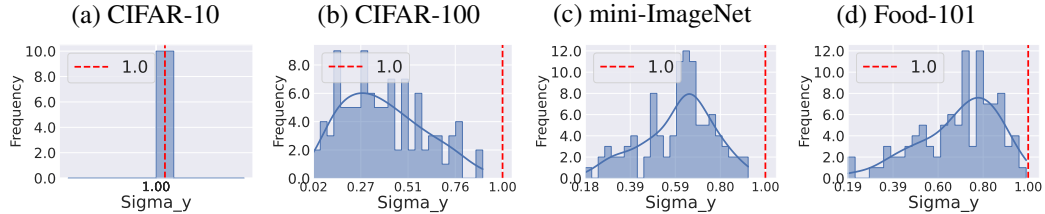
Figure 19: Verification of condition numbers $\{\sigma_y\}_{y=1}^C$ of Equation 6 when epoch $= 200$ and $\alpha = 0.1$ with $\rho = 0.5$ MAJ. Vertical dashed lines represent the value 1, and we observe that all the condition numbers are smaller than 1. This verifies the validity of the condition for Lemma 4.2, and thus confirms that RC3P produces smaller prediction sets than CCP by the optimized trade-off between calibration on non-conformity scores and calibrated label ranks.

35

| (a) CIFAR-10 | (b) CIFAR-100 | (c) mini-ImageNet | (d) Food-101 |
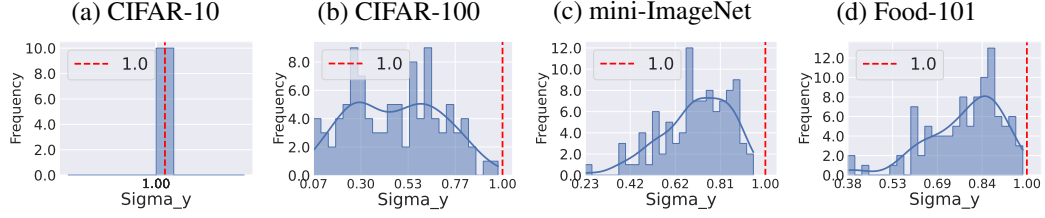
Figure 20: Verification of condition numbers $\{\sigma_y\}_{y=1}^C$ of Equation 6 when epoch $= 50$ and $\alpha = 0.1$ with $\rho = 0.1$ EXP. Vertical dashed lines represent the value 1, and we observe that all the condition numbers are smaller than 1. This verifies the validity of the condition for Lemma 4.2, and thus confirms that RC3P produces smaller prediction sets than CCP by the optimized trade-off between calibration on non-conformity scores and calibrated label ranks.

| (a) CIFAR-10 | (b) CIFAR-100 | (c) mini-ImageNet | (d) Food-101 |
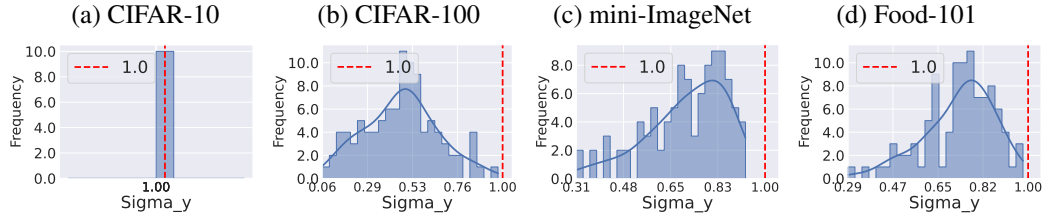
Figure 21: Verification of condition numbers $\{\sigma_y\}_{y=1}^C$ of Equation 6 when epoch $= 50$ and $\alpha = 0.1$ with $\rho = 0.5$ EXP. Vertical dashed lines represent the value 1, and we observe that all the condition numbers are smaller than 1. This verifies the validity of the condition for Lemma 4.2, and thus confirms that RC3P produces smaller prediction sets than CCP by the optimized trade-off between calibration on non-conformity scores and calibrated label ranks.

| (a) CIFAR-10 | (b) CIFAR-100 | (c) mini-ImageNet | (d) Food-101 |
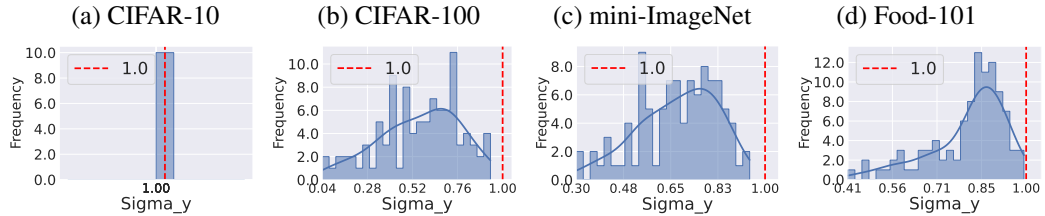
Figure 22: Verification of condition numbers $\{\sigma_y\}_{y=1}^C$ of Equation 6 when epoch $= 50$ and $\alpha = 0.1$ with $\rho = 0.1$ POLY. Vertical dashed lines represent the value 1, and we observe that all the condition numbers are smaller than 1. This verifies the validity of the condition for Lemma 4.2, and thus confirms that RC3P produces smaller prediction sets than CCP by the optimized trade-off between calibration on non-conformity scores and calibrated label ranks.

| (a) CIFAR-10 | (b) CIFAR-100 | (c) mini-ImageNet | (d) Food-101 |
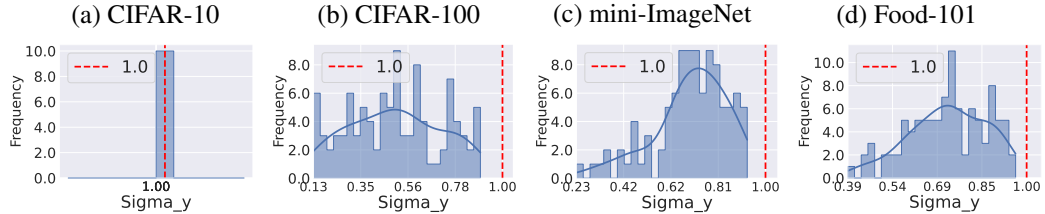
Figure 23: Verification of condition numbers $\{\sigma_y\}_{y=1}^C$ of Equation 6 when epoch $= 50$ and $\alpha = 0.1$ with $\rho = 0.5$ POLY. Vertical dashed lines represent the value 1, and we observe that all the condition numbers are smaller than 1. This verifies the validity of the condition for Lemma 4.2, and thus confirms that RC3P produces smaller prediction sets than CCP by the optimized trade-off between calibration on non-conformity scores and calibrated label ranks.

36

(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101
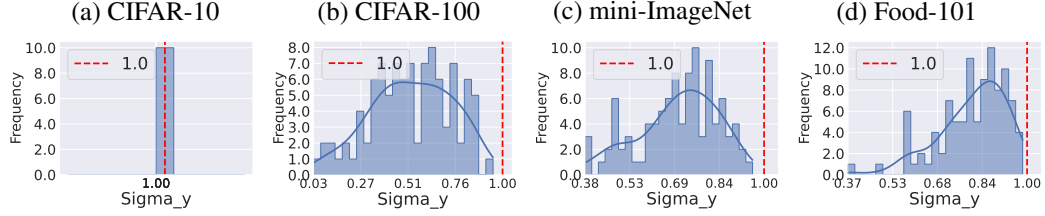
Figure 24: Verification of condition numbers $\{\sigma_y\}_{y=1}^C$ of Equation 6 when epoch $= 50$ and $\alpha = 0.1$ with $\rho = 0.1$ MAJ. Vertical dashed lines represent the value 1, and we observe that all the condition numbers are smaller than 1. This verifies the validity of the condition for Lemma 4.2, and thus confirms that RC3P produces smaller prediction sets than CCP by the optimized trade-off between calibration on non-conformity scores and calibrated label ranks.

(a) CIFAR-10 (b) CIFAR-100 (c) mini-ImageNet (d) Food-101
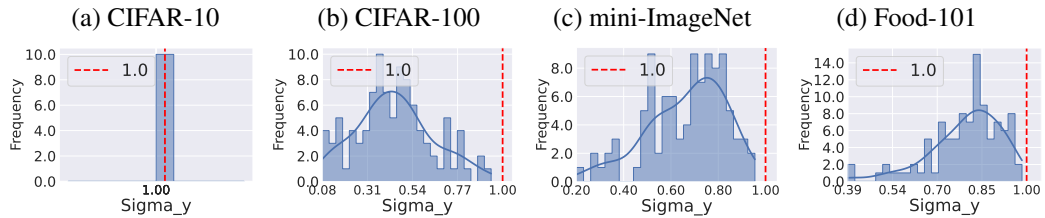
Figure 25: Verification of condition numbers $\{\sigma_y\}_{y=1}^C$ of Equation 6 when epoch $= 50$ and $\alpha = 0.1$ with $\rho = 0.5$ MAJ. Vertical dashed lines represent the value 1, and we observe that all the condition numbers are smaller than 1. This verifies the validity of the condition for Lemma 4.2, and thus confirms that RC3P produces smaller prediction sets than CCP by the optimized trade-off between calibration on non-conformity scores and calibrated label ranks.

37

C.11 Complete Experiment Results on Balanced Classification Datasets

In this subsection, we report complete experimental results over four balanced datasets and $\alpha = 0.1$. Specifically, Figure 26 shows the class-conditional coverage and the corresponding prediction set sizes. From the first row of Fig 26, the class-wise coverage bars of CCP and RC3P distribute on the right-hand side of the target probability $1 - \alpha$ (red dashed line). Second, RC3P outperforms CCP and Cluster-CP with 24.47% (on four datasets) or 32.63% (excluding CIFAR-10) on imbalanced datasets and 32.63% on balanced datasets decrease in terms of average prediction set size the same class-wise coverage. The second row of Figure 26 shows (i) RC3P has more concentrated class-wise coverage distribution than CCP and Cluster-CP; (ii) the distribution of prediction set sizes produced by RC3P is globally smaller than that produced by CCP and Cluster-CP, which is justified by a better trade-off number of $\{\sigma_y\}_{y=1}^{K}$ as shown in Figure 3.

Figure 27 illustrates the normalized frequency distribution of label ranks included in the prediction sets on balanced datasets. It is evident that the distribution of label ranks in the prediction set generated by RC3P tends to be lower compared to those produced by CCP and Cluster-CP. Furthermore, the probability density function tail for label ranks in the RC3P prediction set is notably shorter than that of other methods. This indicates that RC3P more effectively incorporates lower-ranked labels into prediction sets, as a result of its augmented rank calibration scheme.

Figure 28 verifies the condition numbers σ_y on balanced datasets. This result verifies the validity of Lemma 4.2 and Equation 6 and confirm that the optimized trade-off between the coverage with inflated quantile and the constraint with calibrated rank leads to smaller prediction sets.
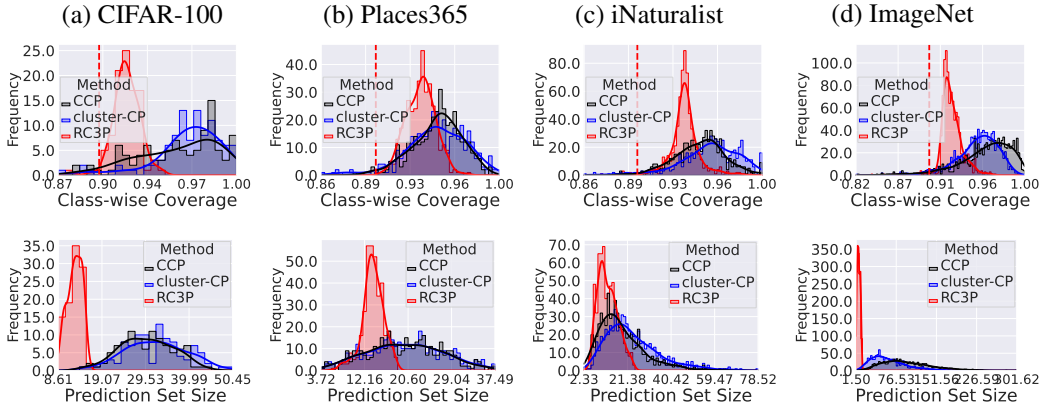
Figure 26: Class-conditional coverage (Top row) and prediction set size (Bottom row) achieved by CCP, Cluster-CP, and RC3P methods when $\alpha = 0.1$ on four balanced datasets. It is clear that RC3P has more densely distributed class-conditional coverage above 0.9 (the target $1 - \alpha$ class-conditional coverage) than CCP and Cluster-CP with significantly smaller prediction sets on all datasets.
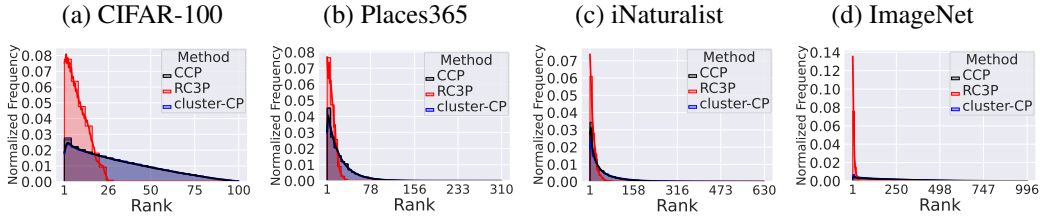
Figure 27: Visualization for the normalized frequency distribution of label ranks included in the prediction set of CCP, Cluster-CP, and RC3P with $\rho = 0.1$ on balanced datasets. It is clear that the distribution of normalized frequency generated by RC3P tends to be lower compared to those produced by CCP and Cluster-CP. Furthermore, the probability density function tail for label ranks in the RC3P prediction set is notably shorter than that of other methods.

38
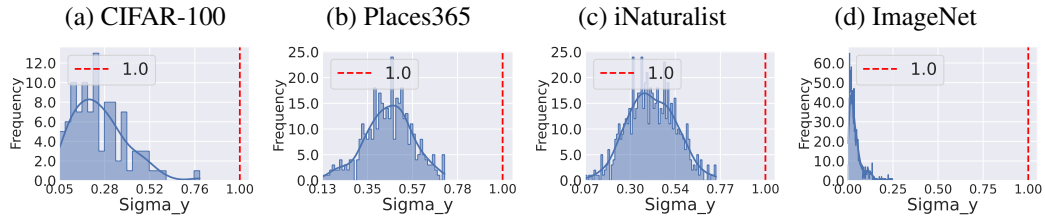
Figure 28: Verification of condition numbers $\{\sigma_y\}_{y=1}^{K}$ in Equation 6 on balanced datasets. Vertical dashed lines represent the value 1, and we observe that all the condition numbers are smaller than 1. This verifies the validity of the condition for Lemma 4.2, and thus confirms that `RC3P` produces smaller prediction sets than `CCP` using calibration on both non-conformity scores and label ranks.

39

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction has stated the contributions and important assumptions of our paper and match our theoretical and experimental results. We have summarize all claims at the end of introduction.

   Guidelines:

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

2. **Limitations**

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

   Answer: [No]

   Justification: A limitation of our paper is that we assume that (X_i, Y_i) are exchangeable (for example, i.i.d.). This assumption is common and fundamental in CP works, so we do not discuss in our paper.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

40

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

Answer: [Yes]

Justification: In Section 4.1 and 4.2, we provide the the full set of assumptions for our theoretical result. Corresponding proofs are provided in Appendix B.

Guidelines:

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

4. **Experimental Result Reproducibility**

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

Answer: [Yes]

Justification: We have provided all all the information needed to reproduce the experiments, including experiments setting, evaluation metric and codes.

Guidelines:

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

41

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

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

   Answer: [Yes]

   Justification: We have provided the codes in the supplementary material.

   Guidelines:

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

6. **Experimental Setting/Details**

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

   Answer: [Yes]

   Justification: We have provided the detail about dataset, training and calibration in Section 5.1 and Appendix C.1, C.2, and C.3.

   Guidelines:

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

7. **Experiment Statistical Significance**

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

   Answer: [Yes]

   Justification: We have provide the standard deviation of our main results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

42

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

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

Answer: [No]

Justification: We follow the training setting of previous papers, so we choose to not discuss the computer resources.

Guidelines:

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

9. **Code Of Ethics**

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

Answer: [Yes]

Justification: Our work strictly adheres to the NeurIPS Code of Ethics.

Guidelines:

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

10. **Broader Impacts**

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

Answer: [No]

Justification: The methodological improvements gained in our paper can lead to improvements in safe deployment of classifiers in human-ML collaborative systems. We do not anticipate any negative ethical or societal impact.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

43

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

Answer: [NA]

Justification: Our experiments are conducted on public and benchmark datasets.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: All the assets used in the paper are open-source and have been properly cited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

44

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

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

    Answer: [Yes]

    Justification: We have provided anonymized zip file in supplementary materiel.

    Guidelines:

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

14. **Crowdsourcing and Research with Human Subjects**

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

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

45