Quantifying and Improving Adaptivity in Conformal Prediction Through Input Transformations

Sooyong Jang[1], **Insup Lee**[1]
[1] PRECISE Center
University of Pennsylvania
{sooyong, lee}@seas.upenn.edu

Abstract

Conformal prediction constructs a set of labels instead of a single point prediction, while providing a probabilistic coverage guarantee. Beyond the coverage guarantee, adaptiveness to example difficulty is an important property. It means that the method should produce larger prediction sets for more difficult examples, and smaller ones for easier examples. Existing evaluation methods for adaptiveness typically analyze coverage rate violation or average set size across bins of examples grouped by difficulty. However, these approaches often suffer from imbalanced binning, which can lead to inaccurate estimates of coverage or set size. To address this issue, we propose a binning method that leverages input transformations to sort examples by difficulty, followed by uniform-mass binning. Building on this binning, we introduce two metrics to better evaluate adaptiveness. These metrics provide more reliable estimates of coverage rate violation and average set size due to balanced binning, leading to more accurate adaptivity assessment. Through experiments, we demonstrate that our proposed metric correlates more strongly with the desired adaptiveness property compared to existing ones. Furthermore, motivated by our findings, we propose a new adaptive prediction set algorithm that groups examples by estimated difficulty and applies group-conditional conformal prediction. This allows us to determine appropriate thresholds for each group. Experimental results on both (a) an Image Classification (ImageNet) (b) a medical task (visual acuity prediction) show that our method outperforms existing approaches according to the new metrics.

Code — https://github.com/sooyongj/t-aps
Extended version — https://arxiv.org/pdf/2511.11472

1 Introduction

Researchers have used conformal prediction to ensure the reliability of machine learning predictions. Unlike traditional models that produce a single point prediction, conformal prediction outputs a set of labels. This set ensures that it contains the true label with high probability. Beyond this coverage guarantee, adaptivity to example difficulty is a crucial property (Seedat et al. 2023). It enables the method to produce smaller prediction sets for easier instances while allocating larger sets to more challenging ones.

A key challenge in developing adaptive prediction sets is the precise evaluation of adaptiveness as existing standard metrics have limitations. A common approach involves comparing the coverage rate violation or the average set size across subgroups of examples, where the subgroups are constructed based on example difficulty. Typically, difficulty is inferred either from the prediction set size or from the rank of the ground truth label within the softmax output. In particular, previous works have measured the worst-case coverage rate violation across bins based on set size (*e.g.,* SSCV (Angelopoulos et al. 2021), ESCV (Huang et al. 2024b)), or have compared changes in coverage rate violation (Angelopoulos et al. 2021) or relative average set size changes (Huang et al. 2024b) across bins based on the ground truth rank.

However, these methods have a notable weakness: the estimation of the coverage rate violation or average set size within each bin can be unreliable due to the imbalance in the number of examples across bins. If a bin contains few examples, the estimation becomes highly biased. Conversely, if a bin contains the majority of examples, the coverage rate violation will be close to zero, since a prediction set algorithm is designed to closely match the target coverage rate.

A straightforward way to mitigate the issue of uneven number of examples in each bin is to apply the uniform-mass binning (Kumar, Liang, and Ma 2019; Gupta, Podkopaev, and Ramdas 2020), where each bin contains approximately the same number of examples. However, it is challenging to apply uniform-mass binning with the ground truth rank (one well-known method of estimating example difficulty (Angelopoulos et al. 2021; Huang et al. 2024b)), due to the long-tailed nature of its distribution. Most examples have a high rank (corresponding to a high softmax value for the ground truth), while only a few examples have a low rank. For example, a classifier with 80 % accuracy assigns rank one to the correct label in 80 % of cases. This skew makes uniform-mass binning difficult to apply, as several bins may end up containing only examples with rank one, resulting in homogeneous groups that do not meaningfully reflect varying difficulty.

Therefore, we propose an alternative sorting method based on a different difficulty estimation. Specifically, we estimate the difficulty $D(x) \in [0, 1]$ (or equivalently, its ease as $E(x) = 1 - D(x)$) by measuring the variability in the model's predictions after applying small perturbations to

the input. For example, if we add a small noise δ to an input x, an easy example would result in similar predictions ($f(x) \approx f(x+\delta)$). We empirically demonstrate that this difficulty-based sorting helps form bins that mitigate the aforementioned issue.

Building on this idea, we propose the two improved metrics to better evaluate the adaptiveness according to the difficulty. Our metrics first construct bins by leveraging the difficulty estimated through input transformations. Then, they compute (1) the worst-case coverage rate violation similar to SSCV and ESCV and (2) the relationship between average set size and the average difficulty (ground truth label ranks) across the bins. The first metric assesses if a prediction set algorithm maintains coverage guarantees across difficulty level. The second metric evaluates whether the algorithm constructs prediction set with appropriate size according to the difficulty. While the first metric is straightforward, as it analyzes the coverage rate violation across bins (consistent with prior work (SSCV and ESCV)), the second metric requires additional analysis to determine its validity as an adaptivity metric. We define the desired adaptiveness property and show that our proposed metric achieves the highest statistical correlation with the property satisfaction rate on the ImageNet (Russakovsky et al. 2015).

Since this experiment shows that we can estimate difficulty using input transformations, we further propose a new adaptive prediction set algorithm by extending this idea. The key extension is to group examples based on estimated difficulty and to separately apply the prediction set algorithm within each group. This separate treatment allows us to learn different thresholds for each group, leading to more appropriate set sizes according to difficulty. To ensure coverage guarantee, we build on ideas from group-conditional conformal prediction (Vovk et al. 2003; Jung et al. 2023; Gibbs, Cherian, and Candès 2025; Jin and Ren 2025), where coverage is enforced not only marginally but also within each predefined subgroup.

We evaluate our proposed prediction set algorithm on two tasks: an image classification task and a visual acuity prediction task. Our experiments demonstrate that the proposed method outperforms existing baselines. Specifically, it achieves lower worst coverage rate violations and establishes a closer relationship between average set size and average example difficulty.

Our contributions are as follows:

- We propose two metrics for evaluating adaptiveness, which provide better assessment compared to existing metrics (Section 3).
- We introduce an adaptive conformal prediction set algorithm that leverages transformation-based binning and group-conditional conformal prediction (Section 4).
- We empirically demonstrate that our proposed method outperforms other baselines across nine base classifiers on ImageNet (Section 5) and four base classifiers on visual acuity prediction (Section 6).

2 Related Work

Adaptiveness Evaluation Metric. Several metrics exist to evaluate adaptivity. SSCV (Angelopoulos et al. 2021) and ESCV (Huang et al. 2024b) assess coverage rate violations across different bins based on prediction set sizes, measuring whether the algorithms maintain the target coverage rate across varying levels of difficulty. Metrics "Deficit" and "Excess" (Navratil, Arnold, and Elder 2021; Seedat et al. 2023)) quantify how tightly the prediction sets are aligned with the ground truth labels. Additionally, previous studies employ empirical analyses, such as analyzing coverage rate violations and average set sizes stratified by ground truth rank (Angelopoulos et al. 2021; Huang et al. 2024b). Section 3 provides a more detailed analysis of these metrics.

Adaptive Prediction Set. Multiple prior works use softmax scores to construct adaptive prediction sets, allowing set sizes to vary according to example difficulty (Sadinle, Lei, and Wasserman 2019; Romano, Sesia, and Candes 2020; Angelopoulos et al. 2021; Huang et al. 2024b). They incorporate softmax outputs into their non-conformity score functions, leveraging either individual softmax scores (Sadinle, Lei, and Wasserman 2019) or the full softmax score distribution over labels (Romano, Sesia, and Candes 2020; Angelopoulos et al. 2021; Huang et al. 2024b) to assess difficulty.

Transformation and Uncertainty Quantification. Earlier studies show that transformations help quantify uncertainty. Ayhan and Berens (2018) compute the interquartile range (IQR) of the top softmax scores after applying input transformations, using it as a measure of uncertainty to identify challenging cases in a simulated clinical workflow. This idea further extends to confidence estimation (calibration) (Bahat and Shakhnarovich 2020; Jang, Lee, and Weimer 2021; Rizve, Kardan, and Shah 2022). Bahat and Shakhnarovich (2020) use the mean of top softmax scores for perturbed images as confidence, while Jang, Lee, and Weimer (2021) leverage lossy label-invariant transformations to group examples for confidence calibration. Similarly, Rizve, Kardan, and Shah (2022) quantify uncertainty by measuring the variance of the top softmax scores across perturbed images, using this value for temperature scaling. In addition, Tao et al. (2024) use prediction consistency for uncertainty estimation. Although Guo et al. (2025) do not use transformations, they estimate uncertainty based on the logit gap.

Group-Conditional Conformal Prediction. Beyond marginal coverage guarantee, a growing body of research focuses on group-conditional coverage guarantee. This setting aims to ensure that prediction sets achieve the target coverage rate within predefined subgroups of examples. For example, in disease prediction, the model should maintain consistent coverage across different demographic groups, such as race. Various researchers propose algorithms for group-conditional conformal prediction (Vovk et al. 2003; Jung et al. 2023; Gibbs, Cherian, and Candès 2025; Jin and Ren 2025), with different properties. These differences include whether the groups can overlap and whether the grouping is based soley on features or other factors.

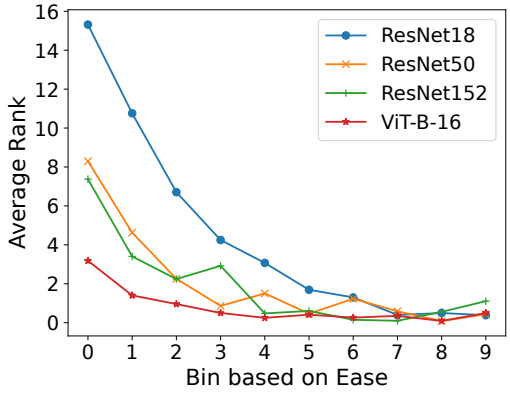


Figure 1: Ease vs. Average Rank.

3 Transformation Based Metric

This section describes metrics for evaluating adaptivity. Navratil, Arnold, and Elder (2021); Seedat et al. (2023) introduce "Deficit" and "Excess" (Navratil, Arnold, and Elder 2021; Seedat et al. 2023) for regression problems. Specifically, Deficit quantifies the number of classes required to include the ground truth, whereas Excess assess the number of classes that can be removed while maintaining coverage, both following softmax score order. SSCV (Angelopoulos et al. 2021), a widely used metrics, quantifies worst-case deviation by comparing the target coverage rate with the empirical coverage rate across different size strata. Similarly, ESCV (Huang et al. 2024b) relates closely to SSCV but considers individual prediction set sizes rather than size ranges. Additionally, two empirical analyses (Rank-CV, Rank-SS) utilize ground truth rank-based binning. The ground truth rank refers to the position of the true label within the sorted softmax scores. Using this binning, Rank-CV compares coverage violation across bins to assess consistency, while Rank-SS examines whether average set sizes increase with rising average ground truth rank.

However, these metrics have certain limitations. Deficit and Excess are biased toward one side; for example, prediction sets containing all labels achieve zero Deficit. Both SSCV and ESCV use set size-based binning, which can yield inaccurate metric when bins have few examples. Rank-CV and Rank-SS share SSCV and ESCV's limitation due to their similar binning strategy.

We focus on a common limitation of SSCV, ESCV, and the two empirical analysis methods: all rely on binning examples. Such methods may yield inaccurate estimates when bins are imbalanced. To address this, we propose using a uniform-mass binning, which ensures each bin contains approximately the same number of examples. However, applying uniform-mass binning based on ground truth label rank is challenging, due to its long-tailed distributions, *i.e.,* most examples have a rank of one. Thus, a more suitable difficulty sorting method is required to ensure that the average difficulty across bins varies smoothly.

We leverage a method to estimate example difficulty by utilizing input transformations. At a high level, if the output

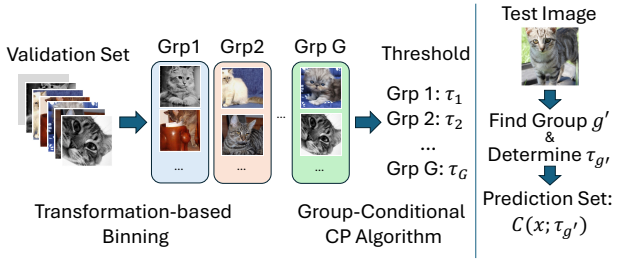


Figure 2: Algorithm Process. The left part illustrates the calibration process, while the right part shows the construction of prediction sets for test images.

of a prediction model f remains unchanged after applying an input transformation, such as adding small noise to an example x, the example x can be considered easy for the model. Accordingly, we define the ease of example x with respect to the model f, denoted as $E : \mathcal{X} \to [0, 1]$, formally:

$$E(x; f) = \frac{1}{L} \sum_{l=1}^{L} \text{sim}(f(x), f(t_l(x))), \tag{1}$$

where $\text{sim}(x, y)$ denotes a similarity between x and y, such as the cosine similarity, and $t_l(x)$ represents a transformation (*e.g.,* $t_l(x) = x + \delta_l$, where $\delta_l \sim \mathcal{N}(0, \sigma^2)$).

Regarding the input transformations, we use Gaussian noise for simplicity; other transformations (*e.g.,* color jitters) work if severity is controlled — excessive change mostly alters predictions. Adversarial perturbations are less suitable, but may work with robust models.

To verify the effectiveness of this sorting's method, we plot the average ground truth rank within each bin on ImageNet dataset using the four different base classifiers. As shown in Figure 1, for all base classifiers, the average rank generally decreases as the ease increases, confirming the validity of our sorting strategy for rank-based binning.

We now describe how examples are binned using $E(x; f)$. We partition the range of ease scores $[0, 1]$ into B bins using uniform-mass binning. Formally, let $a : \{1, ..., N\} \to \{1, ..., N\}$ be an indexing function that sorts examples such that $E(x_{a(1)}) \leq E(x_{a(2)}) \leq \cdots \leq E(x_{a(N)})$. Then, the set of indices for bin b is defined as:

$$\mathcal{I}_b = \left\{ a(i) \middle| \left\lfloor \frac{(b-1)N}{B} \right\rfloor < i \leq \left\lfloor \frac{bN}{B} \right\rfloor \right\} \quad \text{for } b = 1, 2, \ldots, B. \tag{2}$$

This process is summarized in Algorithm 1 in the appendix.

Based on this transformation-based binning, we define two evaluation metrics. (1) Transformation-Based binning Coverage Violation (T-CV): the worst-case coverage rate violation across bins, and (2) Transformation-Based binning Set Size relationship (T-SS): the relationship between the average set size and the difficulty across bins. The two metrics are formally defined as:

$$\textbf{T-CV} = \max_{b \in [B]} \left| \frac{\sum_{i \in \mathcal{I}_b} \mathbb{I}(y_i \in C(x_i))}{|\mathcal{I}_b|} - (1 - \alpha) \right| \tag{3}$$

Rank	Defict	Excess	SSCV	ESCV	T-SS
1	12	0	0	0	24
2	23	0	1	0	12
3	0	9	16	11	0
4	1	12	11	12	0
5	0	15	8	13	0
Avg.	1.72	4.17	3.58	4.07	**1.33**

Table 1: Metric comparison result. We rank each metric across thirty six cases, based on nine models and four algorithms, with its frequency and average rank reported.

$$\textbf{T-SS} = R^2(\{r(\mathcal{I}_1), \cdots, r(\mathcal{I}_B)\}, \{s(\mathcal{I}_1), \cdots, s(\mathcal{I}_B)\}), \tag{4}$$

where R^2 is the (signed) R-squared value (Coefficient of Determination) and $r(\mathcal{I}_b), s(\mathcal{I}_b)$ are the average rank and average set size for the group of examples $\mathcal{I}_b$, respectively. Here, $R^2(X, Y) = \text{sign}(a) \times \max\left(0, 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}\right)$, where a is the coefficient for a linear regressor, $y, \hat{y}, \bar{y}$ are the true value, predicted value, and mean true value, respectively. Thus, T-SS measures how closely the average rank and average set size of each bin are linearly related, with the sign indicating the direction of their relationship (positive or negative correlation).

While T-CV directly evaluates coverage rate violation, similar to prior metrics (SSCV and ESCV), T-SS focuses on the different aspect, average set size. Therefore, the effectiveness of T-SS needs to be further validated as an adaptivity metric. To validate T-SS, we define a desired property (Property 1) that adaptive prediction sets should satisfy and analyze the correlation between each metric and the degree of this property being met. The next subsection describes the experimental setup and presents the results.

Property 1. *(Difficulty and Set Size Relationship). For any two subsets g_1, g_2 of size m, let d_1, d_2 be their average difficulty and s_1, s_2 be their average prediction set size. Then, a desirable adaptivity property requires that one of the following holds:*

$$d_1 \leq d_2 \implies s_1 \leq s_2 \quad \textit{or} \quad d_1 > d_2 \implies s_1 > s_2.$$

In words, groups of harder examples should correspond to larger average prediction set size.

Experiment. We conduct an experiment to compare T-SS (Equation 4) with existing metrics, focusing on their ability to estimate the satisfaction rate of the desired property. We perform $R = 1000$ trials, and in each trial, we sample $M = 2000$ examples from ImageNet validation set. For each M examples, we estimate the property satisfaction rate by sampling $m = 100$ examples $T = 10000$ times and checking whether the property holds. We also compute each metric on the same M examples. Finally, we calculate the Spearman correlation (Spearman 1904) between the property satisfaction rates and the corresponding metric values across the R trials.

We use nine base classifiers (ResNet18, ResNet50, ResNet152 (He et al. 2016), ViT-B-16, ViT-L-16, ViT-H-14 (Dosovitskiy et al. 2020), EfficientNet-V2-M, EfficientNet-V2-L (Tan and Le 2021), Swin-V2-B (Liu et al. 2022)) and four prediction set algorithms (LAC (Sadinle, Lei, and Wasserman 2019), APS (Romano, Sesia, and Candes 2020), RAPS (Angelopoulos et al. 2021), SAPS (Huang et al. 2024b)) on ImageNet.

We rank the five metrics based on the Spearman correlation coefficients and summarize the result in Table 1. We provide the full results, including p-values in Table 3 in the appendix. As shown in Table 1, T-SS outperforms existing metrics, achieving the best rank in most cases. Moreover, the correlation coefficients for T-SS are statistically significant based on the corresponding p-values. This implies that T-SS is strongly correlated with the property satisfaction rate, indicating its effectiveness in evaluating adaptivity.

4 Prediction Set Algorithm

In this section, we describe our algorithm, which is based on the transformation-based binning introduced in Section 3.

4.1 Algorithm Overview

As illustrated in Section 3, transformation-based binning effectively groups examples by similar difficulty. Building on this, we argue that applying conformal prediction separately within each group enables the algorithm to determine more appropriate thresholds, leading to improved adaptivity. To ensure coverage guarantees in this group-wise setting, we adopt an idea from group-conditional conformal prediction.

The group conditional coverage guarantee was proposed to address the limitations of marginal coverage. Instead of the marginal coverage guarantee, the group-conditional coverage guarantee ensures:

$$\mathbb{P}_{(x_i,y_i)\sim\mathcal{P}}[y_i \in C(x_i) | G(x_i, y_i) = g] \geq 1-\alpha, \quad \forall g \in [H], \tag{5}$$

where $\mathcal{P}$ is the data distribution, $C(x_i)$ is the prediction set, $G(x_i, y_i)$ is the group assignment function, α is the target miscoverage rate, and H is the number of groups.

We define the group assignment function $G(x_i, y_i) := b$ such that $i \in \mathcal{I}_b$, and apply a group-conditional conformal prediction algorithm based on this assignment.

4.2 Algorithm Detail

The algorithm is described in Figure 2 and Algorithm 2 (Appendix). At a high level, we first group examples based on their estimated difficulty using transformation-based binning (Left part in Figure 2) and then apply a group-conditional conformal prediction algorithm to obtain different conformal predictors for each group (Middle). Given a test data point x, we identify its corresponding bin, and use the threshold assigned to that bin to construct the prediction set (Right).

Multiple group-conditional conformal algorithms (Vovk et al. 2003; Jung et al. 2023; Ding et al. 2023; Gibbs, Cherian, and Candès 2025; Jin and Ren 2025) exist for $\mathcal{A}$. These algorithms differ in aspects such as whether the group overlapping is allowed (Jung et al. 2023; Gibbs, Cherian, and Candès 2025; Jin and Ren 2025) and whether the group assignment depends solely based on the feature x or on both the feature x (Jung et al. 2023; Gibbs, Cherian, and Candès 2025;

Base Model	Acc.	T-CV						T-SS					
		LAC	APS	RAPS	SAPS	O-LAC	O-SAPS	LAC	APS	RAPS	SAPS	O-LAC	O-SAPS
$\alpha = 0.10$													
ResNet18	69.89	0.192	**0.040**	0.055	0.255	0.060	0.053	0.869	0.924	0.921	0.827	**0.926**	0.912
ResNet50	80.81	0.323	0.148	0.212	0.290	0.068	**0.057**	0.512	0.210	0.617	0.620	**0.835**	0.807
ResNet152	82.26	0.243	0.060	0.273	0.260	0.052	**0.047**	0.674	0.756	0.614	0.728	**0.783**	0.769
ViT-B-16	85.22	0.297	0.065	**0.045**	0.055	0.050	**0.045**	0.636	**0.814**	0.810	0.743	0.797	0.755
ViT-L-16	88.20	0.463	0.075	0.180	**0.040**	0.042	**0.040**	0.296	**0.823**	0.558	0.760	0.793	0.776
ViT-H-14	88.53	0.430	0.043	0.160	0.050	0.043	**0.040**	0.287	0.611	0.495	0.619	**0.627**	0.623
Efficientnet-V2-M	85.11	0.258	0.050	0.230	0.253	0.045	**0.040**	0.517	0.408	0.368	0.554	**0.624**	0.595
Efficientnet-V2-L	85.68	0.290	0.050	0.190	0.057	0.050	**0.045**	0.473	0.507	0.479	0.538	**0.557**	0.553
Swin-V2-B	83.93	0.243	0.050	0.050	0.075	**0.040**	**0.040**	0.627	**0.703**	0.686	0.670	0.694	0.694
Average	-	0.304	0.065	0.155	0.148	0.050	**0.045**	0.543	0.640	0.616	0.673	**0.737**	0.720
Avg. Rank from CD	-	5.44	2.44	3.89	4.11	1.33	**1.11**	5.33	2.67	3.56	3.56	**1.11**	1.67
$\alpha = 0.05$													
ResNet18	69.89	0.112	**0.033**	0.045	0.165	**0.033**	0.035	0.912	0.926	0.920	0.828	**0.930**	0.910
ResNet50	80.81	0.153	0.110	0.165	0.100	**0.035**	0.040	0.680	0.095	0.474	0.636	**0.828**	0.807
ResNet152	82.26	0.110	0.033	0.163	0.107	**0.030**	0.033	0.763	0.689	0.679	0.727	**0.791**	0.769
ViT-B-16	85.22	0.107	0.043	0.033	0.053	**0.030**	0.033	0.700	**0.802**	0.770	0.725	**0.802**	0.745
ViT-L-16	88.20	0.143	0.040	0.090	0.058	**0.030**	**0.030**	0.690	0.803	0.711	0.743	**0.813**	0.776
ViT-H-14	88.53	0.115	0.030	**0.028**	0.048	0.030	**0.028**	0.593	0.542	0.585	0.613	**0.621**	0.619
Efficientnet-V2-M	85.11	0.110	**0.030**	0.145	0.112	0.030	**0.028**	0.558	0.402	0.504	0.554	**0.627**	0.598
Efficientnet-V2-L	85.68	0.110	0.035	0.115	0.080	**0.033**	**0.033**	0.536	0.478	0.491	0.536	**0.570**	0.552
Swin-V2-B	83.93	0.107	0.035	0.050	0.055	**0.028**	0.030	0.666	0.690	0.678	0.670	**0.705**	0.695
Average	-	0.119	0.043	0.093	0.086	**0.031**	0.032	0.678	0.603	0.646	0.670	**0.743**	0.719
Avg. Rank from CD	-	5.00	1.78	4.11	4.33	**1.00**	**1.00**	3.78	3.67	4.22	3.67	**1.00**	1.89

Table 2: Our Metrics. The second-to-last row shows the average value over different models, and the last row shows the average rank based on Critical Difference (CD) analysis. $B = 50$.

Jin and Ren 2025) and label y (Vovk et al. 2003; Ding et al. 2023). However, in our setting, Mondrian Conformal Prediction (MCP) (Vovk et al. 2003) is sufficient. The approach forms groups and applies conformal prediction individually within each group. Regarding the randomness in the grouping function, separating the calibration sets for binning and for CP preserves group-conditional coverage, whereas using a single calibration set suffices for marginal coverage.

5 Application to Image Classification

We conduct an experiment to compare our proposed algorithm with other adaptive prediction set algorithms.

Dataset and Set-Up. We use ImageNet (Russakovsky et al. 2015) in this section. The original dataset consists of training, validation and test set, but we use only the validation set.

We repeat each experiment $R = 100$ times and we report the median value of each metric as the final result. In each repetition, we randomly sample a validation set D_{val} of size $n_{val} = 20,000$ and a test set D_{test} of size $n_{test} = 20,000$ from the original validation set. For algorithms requiring hyperparameter tuning (including ours), we evenly split D_{val} into a calibration set D_{cal} and a tuning set D_{tune}. Algorithms that do not require hyperparameter tuning use the entire D_{val}. This setup ensures that all methods use the same amount of validation data.

Baselines and Our Algorithm. We compare our method against four adaptive prediction set algorithms: LAC (Sadinle, Lei, and Wasserman 2019), APS (Romano, Sesia, and Candes 2020), RAPS (Angelopoulos et al. 2021), SAPS (Huang et al. 2024b). We implement these baselines using TorchCP library (Huang et al. 2024a). For our method, we use the nonconformity score functions of LAC and SAPS as the base conformal prediction algorithms, and denote the resulting method as O-LAC and O-SAPS, respectively.

Some algorithms, including ours, require tuning hyperparameters, such as the number of bins and the transformations types in our method, the regularization weight in RAPS, and the ranking weight in SAPS. While RAPS uses SSCV and average set size for tuning, we instead use our proposed metric (T-SS) for hyperparameter selection to better align with our goal of improving adaptivity.

For all methods, we set target miscoverage rate $\alpha \in \{0.10, 0.05\}$.

Metrics. We primarily use the two proposed metrics (T-CV and T-SS with the number of bins $B = 50$) to compare the adaptivity of prediction set algorithms. We also compute the empirical coverage rate and average set size to evaluate the coverage guarantee and the efficiency of the algorithms, and include them (Table 4) in the appendix. Further results using existing evaluation metrics are presented in Tables 6 and 7 in the appendix. We also report the performance of our metric with an alternative bin count ($B \in \{100, 30\}$), provided in

the appendix (Tables 8 and 9).

5.1 Results

Table 4 in the appendix shows coverage rate and average set size. Overall, all algorithms achieve empirical coverage rates close to the targets ($1 - \alpha \in \{0.90, 0.95\}$). In terms of average set size, LAC produces the smallest sets on average, while APS yields the largest sets, with the others showing similar sizes. Specifically, binning in our algorithm reduces the samples per bin, making threshold estimation harder and slightly increasing size. However, ours shows comparable sizes to others except LAC (worse adaptivity one).

Table 2 presents the result using our proposed metrics, T-CV and T-SS. The 'Average' row indicates the mean values of the respective metrics, while 'Avg. Rank from CD' represents the average rank of each algorithm derived from the Critical Difference (CD) analysis. For the CD analysis, we first perform Friedman Test, which indicates statistically significant differences among the algorithms (all p-values are significant, < 0.001). We then construct CD diagrams, shown in the appendix (Figures 7 - 14), and rank algorithms based on the CD results, assigning the same rank to algorithms that are not statistically different. Finally, we compute the average rank of each algorithm to compare their relative performance.

Based on the metric values and the average ranks from the CD analysis, we analyze the adaptiveness performance of individual algorithms in detail. First, O-SAPS achieves the smallest T-CV value, with O-LAC achieving the second-best performance. This is noteworthy as there is a general trend where the average set size correlates with T-CV: algorithms with smaller average set sizes (*e.g.*, LAC) tend to have higher T-CV values, while those with larger set sizes (*e.g.*, APS) have lower T-CV value. Given that APS produces the largest average set size, it is expected that APS achieves a low T-CV value. In this context, it is notable that our algorithms record the low T-CV values while maintaining not overly large average set sizes unlike APS.

Regarding T-SS, our two algorithms achieve the best (O-LAC) and the second best (O-SAPS), indicating that the prediction set sizes respond adaptively to example difficulty. Although APS generally performs well in T-SS, it shows a notable exception on ResNet50, where its exceptionally large average set size leads to a weaker relationship between set size and difficulty.

In summary, our methods better adapt the prediction set size according to the difficulty (as indicated by T-SS), while maintaining the coverage rates close to the target level across different difficulty levels (T-CV).

Limitations. One limitation of our approach is that as the base classifier becomes more accurate, the T-SS score tends to decrease. For example, O-LAC achieves a T-SS of 0.926 with ResNet18 and 0.627 with ViT-H-14, even though ViT-H-14 attains a significantly higher classification accuracy (69.89% vs. 88.20%). We hypothesize that this phenomenon is due to the distribution of the ground truth rank. T-SS measures the relationship between the average ground truth rank and the average set size across bins. A more accurate classifier tends to have more examples with the ground truth rank of one. This leads to many bins having similar average ranks, which reduces variability of difficulty and limits maximum T-SS score. In the extreme case of a base classifier achieving 100% accuracy, all bins would have an average rank of one, making it impossible to compute T-SS meaningfully. For a similar reason, our metrics become less effective when the target miscoverage rate α is high, *e.g.,* $\alpha \in \{0.15, 0.20\}$. Please refer to the results provided in the appendix (Table 5). A higher α allows more miscoverage cases, leading to smaller prediction sets overall, including for more difficult examples. This reduces the contrast in set sizes across different difficulty levels, making it harder to accurately assess the relationship between example difficulty and prediction set size — the core aspect that T-SS aims to measure.

Despite these limitations, T-SS remains useful for comparing different prediction set algorithms under a fixed base classifier and the reasonable target miscoverage rate considering the safety. Under these conditions, all methods operate under the same conditions, ensuring a fair and meaningful evaluation.

6 Application to Visual Acuity Prediction

We apply our adaptive prediction set algorithm to a medical task as a practical application. Specifically, we consider visual acuity (VA) prediction, a clinically relevant task studied in prior works (Kim et al. 2022; Paul et al. 2023; Jang et al. 2025). Given a retinal fundus image, x_i, the goal is to predict the visual acuity y_i. The task can be formulated in various ways; here, we adopt the same formulation as in (Jang et al. 2025). In particular, the task is to train a model $f(x)$ for Gaussian distribution over visual acuity (*i.e.,* a regression task), and then apply conformal prediction to obtain prediction intervals with a coverage guarantee.

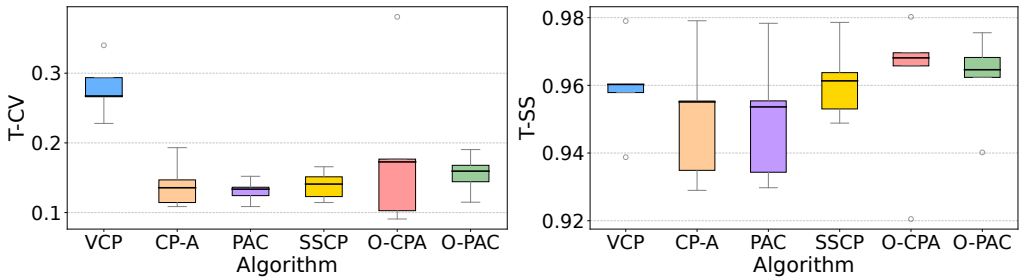
We use the dataset from (Kim et al. 2022) and the trained models from (Jang et al. 2025), and apply our algorithms to them. Because our algorithm is originally designed for classification, we modify it to handle regression. First, we replace cosine similarity with a new similarity metric defined as, $\text{similarity}(y_1, y_2) = \exp(-|y_1 - y_2|)$. Second, instead of using the prediction set size, we evaluate our method using the width of prediction interval.

6.1 Experiment Set-Up

Dataset. We use the same dataset as in (Kim et al. 2022; Jang et al. 2025). For a detailed description of the dataset, please refer to (Jang et al. 2025).

Baselines. We implement four baselines:

- **CP** and **CP-A** are the vanilla CP approaches using different non-conformity score functions: $s_{\text{CP}}(x, y) = |f_\mu(x) - y|$ for CP, and $s_{\text{CP-A}}(x, y) = \frac{|f_\mu(x)-y|}{f_\sigma(x)}$ for CP-A. In other words, CP-A is the extension of CP that incorporates the estimated standard deviation, allowing for prediction interval with varying width.
- **PAC** is the PAC prediction interval algorithm used in the paper (Jang et al. 2025).



(a) T-CV (b) T-SS

Figure 3: Visual Acuity Prediction Result. EfficientNet-V2-S. $\alpha = 0.30$

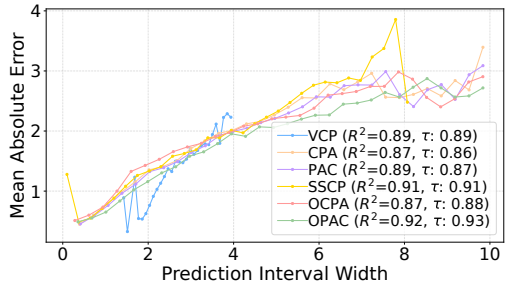


Figure 4: Prediction Interval Width and Prediction Error. EfficientNet-V2-S. $\alpha = 0.30$.

- **SSCP** (Seedat et al. 2023) leverages the loss from a self-supervised learning task loss to estimate uncertainty, which is then used to compute non-conformity score.

We also implement two versions of the adaptive prediction set algorithm, O-CPA and O-PAC which are based on CP-A and PAC, respectively. We use the same model from (Jang et al. 2025), and apply our two versions to each base model.

6.2 Results

Following the set-up in (Jang et al. 2025), we repeat each experiment with five different random seeds. All experiments use a target miscoverage rate $\alpha \in \{0.20, 0.30, 0.40\}$ and for PAC-based methods, we use a significance level of $\delta = 0.00001$. Figure 3 shows results with EfficientNet-V2-S at miscoverage rate $\alpha = 0.30$. Full results appear in the appendix (Tables 10-11, Figures 31-33).

As shown in Figure 3, our methods have similar T-CV values compared to other baselines except VCP, and outperform all baselines in terms of T-SS. Our methods tend to generate conservative intervals when the number of examples is small due to the group-coverage algorithm. This results in an over-coverage rate as shown in Table 11, and a higher T-CV — the over-coverage increases the T-CV value as well. In terms of T-SS, ours achieves higher values, indicating a stronger linear relationships between average error and average interval width, which implies better adaptivity.

We further analyze how effectively each algorithm captures adaptivity by comparing the relationship between prediction interval width and prediction error. The common assumption is that wider intervals are indicative of higher prediction errors, as they reflect greater uncertainty in the base model's predictions. To quantify this relationship, we plot the average prediction interval widths against the mean absolute error. We compute the R^2 value and Kendall's Tau (Kendall 1938) to measure the strength of the relationships between interval width and prediction error. The results for EfficientNet-V2-S with desired miscoverage level of $\alpha = 0.30$ are presented in Figure 4. Our approach achieves higher R^2 and tau values, indicating a stronger positive correlation between interval width and prediction error. This suggests that the prediction intervals constructed by our method more accurately reflect the model's uncertainty, enabling clinicians to utilize interval width as a meaningful measure of uncertainty in predictions. We also note that our methods consistently outperform other algorithms across different base models and target miscoverage levels (Figure 34 in the appendix). Furthermore, for narrow intervals (which are clinically useful), our methods tend to maintain a positive correlation between interval width and prediction error.

7 Conclusion

We identify that existing metrics for evaluating the adaptiveness of conformal prediction sets face challenges primarily due to imbalanced binning. To address this issue, we propose a transformation-based binning method and introduced two new metrics leveraging this approach. We demonstrate that these metrics effectively evaluate the adaptivity; in particular, one of them is effective at measuring the extent to which the desired property for adaptiveness is satisfied. Building on this foundation, we propose an adaptive prediction set algorithm that combines the transformation-based binning with group-conditional conformal prediction. Through experiments on two tasks, we show that our algorithm outperforms existing baselines in terms of the proposed metrics.

Our method builds on the theoretical coverage guarantees of conformal prediction (CP) and group-conditional CP. Extending the theory toward expected set size (Dhillon, Deligiannidis, and Rainforth 2024; Huang et al. 2024b) or establishing tighter bounds connecting the number of samples per bin to calibration (Kumar, Liang, and Ma 2019; Gupta, Podkopaev, and Ramdas 2020) would be a valuable direction for future work.

Acknowledgements

This work was supported by NIH 1R01EY037101 and ARO MURI W911NF-20-1-0080. We gratefully acknowledge Professor Jin Hyun Kim and Professor Yong-Seop Han for providing the visual acuity dataset.

References

Angelopoulos, A. N.; Bates, S.; Jordan, M.; and Malik, J. 2021. Uncertainty Sets for Image Classifiers using Conformal Prediction. In *International Conference on Learning Representations*.

Ayhan, M. S.; and Berens, P. 2018. Test-time data augmentation for estimation of heteroscedastic aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning*.

Bahat, Y.; and Shakhnarovich, G. 2020. Classification Confidence Estimation with Test-Time Data-Augmentation. arXiv:2006.16705.

Dhillon, G. S.; Deligiannidis, G.; and Rainforth, T. 2024. On the expected size of conformal prediction sets. In *International Conference on Artificial Intelligence and Statistics*, 1549–1557. PMLR.

Ding, T.; Angelopoulos, A.; Bates, S.; Jordan, M.; and Tibshirani, R. J. 2023. Class-conditional conformal prediction with many classes. *Advances in neural information processing systems*, 36: 64555–64576.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gibbs, I.; Cherian, J. J.; and Candès, E. J. 2025. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkaf008.

Guo, H.; Tao, L.; Luo, H.; Dong, M.; and Xu, C. 2025. Sample Margin-Aware Recalibration of Temperature Scaling. arXiv:2506.23492.

Gupta, C.; Podkopaev, A.; and Ramdas, A. 2020. Distribution-free binary classification: prediction sets, confidence intervals and calibration. *Advances in Neural Information Processing Systems*, 33: 3711–3723.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, J.; Song, J.; Zhou, X.; Jing, B.; and Wei, H. 2024a. TorchCP: A Python Library for Conformal Prediction. arXiv:2402.12683.

Huang, J.; Xi, H.; Zhang, L.; Yao, H.; Qiu, Y.; and Wei, H. 2024b. Conformal Prediction for Deep Classifier via Label Ranking. In *International Conference on Machine Learning*, 20331–20347. PMLR.

Jang, S.; Jang, K. J.; Choi, H.; Han, Y.-S.; Lee, S.; Kim, J.-h.; and Lee, I. 2025. Fundus Image-based Visual Acuity Assessment with PAC-Guarantees. In Hegselmann, S.; Zhou, H.; Healey, E.; Chang, T.; Ellington, C.; Mhasawade, V.; Tonekaboni, S.; Argaw, P.; and Zhang, H., eds., *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, 535–549. PMLR.

Jang, S.; Lee, I.; and Weimer, J. 2021. Improving classifier confidence using lossy label-invariant transformations. In *International Conference on Artificial Intelligence and Statistics*, 4051–4059. PMLR.

Jin, Y.; and Ren, Z. 2025. Confidence on the focal: Conformal prediction with selection-conditional coverage. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkaf016.

Jung, C.; Noarov, G.; Ramalingam, R.; and Roth, A. 2023. Batch Multivalid Conformal Prediction. In *International Conference on Learning Representations (ICLR)*.

Kendall, M. G. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2): 81–93.

Kim, J. H.; Jo, E.; Ryu, S.; Nam, S.; Song, S.; Han, Y. S.; Kang, T. S.; Lee, W.; Lee, S.; Kim, K. H.; et al. 2022. A deep learning ensemble method to visual acuity measurement using fundus images. *Applied Sciences*, 12(6): 3190.

Kumar, A.; Liang, P. S.; and Ma, T. 2019. Verified uncertainty calibration. *Advances in neural information processing systems*, 32.

Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. 2022. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12009–12019.

Navratil, J.; Arnold, M.; and Elder, B. 2021. Uncertainty Prediction for Deep Sequential Regression Using Meta Models. arXiv:2007.01350.

Paul, W.; Burlina, P.; Mocharla, R.; Joshi, N.; Li, Z.; Gu, S.; Nanegrungsunk, O.; Lin, K.; Bressler, S. B.; Cai, C. X.; et al. 2023. Accuracy of artificial intelligence in estimating best-corrected visual acuity from fundus photographs in eyes with diabetic macular edema. *JAMA ophthalmology*, 141(7): 677–685.

Rizve, M. N.; Kardan, N.; and Shah, M. 2022. Towards realistic semi-supervised learning. In *European conference on computer vision*, 437–455. Springer.

Romano, Y.; Sesia, M.; and Candes, E. 2020. Classification with valid and adaptive coverage. *Advances in neural information processing systems*, 33: 3581–3591.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Sadinle, M.; Lei, J.; and Wasserman, L. 2019. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525): 223–234.

Seedat, N.; Jeffares, A.; Imrie, F.; and van der Schaar, M. 2023. Improving adaptive conformal prediction using self-supervised learning. In *International Conference on Artificial Intelligence and Statistics*, 10160–10177. PMLR.

Spearman, C. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1): 72–101.

Tan, M.; and Le, Q. 2021. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, 10096–10106. PMLR.

Tao, L.; Guo, H.; Dong, M.; and Xu, C. 2024. Consistency Calibration: Improving Uncertainty Calibration via Consistency among Perturbed Neighbors. arXiv:2410.12295.

Vovk, V.; Lindsay, D.; Nouretdinov, I.; and Gammerman, A. 2003. Mondrian confidence machine. *Technical Report*.