# 1 Introduction

In recent years, the large-scale vision-language model (VLM) has shown impressive advances in a wide range of downstream tasks over the last few years. However, similar to most neural models, VLM lacks the ability of uncertainty qualification, thereby raising concerns about their application in life-threatening tasks such as medical image identification [10] and autonomous driving [16]. Conformal Prediction [2, 3, 13, 14], as a trustworthy machine learning technique, addresses this limitation by offering a prediction set that is guaranteed to contain the ground-true label with high probability (e.g. 95%), thus rendering any pre-trained model uncertainty-aware.

In the literature, while conformal prediction methods have been employed in various domains including classification [12, 1], regression [9, 8], large language model [7, 11], its application in VLM has been largely unexplored. In this report, we investigate how conformal prediction performs under the setting of VLM zero-shot, focusing on three key questions: (1) **will prediction sets violate coverage rate under the zero-shot setting?** (2) **how do different vision backbones affect prediction sets?** (1) **how does the size of the calibration dataset influence prediction sets?**

# 2 Preliminaries

## 2.1 Problem setup

We consider a standard supervised classification problem where the input space and the label space with $K$ classes are denoted as $\mathcal{X}$ and $\mathcal{Y} := \{1, 2, \cdots, K\}$, respectively. We suppose a data set with $N$ samples is given as $\mathcal{I} := \{\boldsymbol{x}_i, y_i\}_{i=1}^N$, sampled $i.i.d$ from a joint data distribution $\mathcal{P}_{\mathcal{X}\mathcal{Y}}$. Let $f : \mathcal{X} \to \mathbb{R}^K$ denote a classifier, which maps an input to an output space. Let $(\boldsymbol{x}, y) \sim \mathcal{P}_{\mathcal{X}\mathcal{Y}}$ be a random data pair and $\boldsymbol{f}_y(\boldsymbol{x})$ denotes the $y$-th element of logits vector $\boldsymbol{f}(\boldsymbol{x})$ corresponding to the ground-truth label $y$. The conditional probability of class $y$ can be approximated by softmax probability $\pi_y(\boldsymbol{x})$, where

$$\pi_y(\boldsymbol{x}) = \sigma(f(\boldsymbol{x}))_y = \frac{e^{f_y(\boldsymbol{x})}}{\sum_{i=1}^K e^{f_i(\boldsymbol{x})}}.$$

where $\sigma$ is a softmax function. Let the softmax probability distribution be

$$\boldsymbol{\pi}(\boldsymbol{x}) = (\pi_1(\boldsymbol{x}), \pi_2(\boldsymbol{x}), \cdots, \pi_K(\boldsymbol{x}))$$

Then, the classification prediction is given by $\hat{y} = \text{argmax}_{y \in \mathcal{Y}} \pi_y(\boldsymbol{x})$.

## 2.2 Conformal prediction

In conformal prediction, we aim to construct a set-valued function $\mathcal{C} : \boldsymbol{x} \to 2^{\mathcal{Y}}$ that maps any input $\boldsymbol{x}$ to a set of possible labels $\mathcal{C}(\boldsymbol{x}) \subset \mathcal{Y}$. Specifically, the set function that contains

the true label at a user-specified error rate $\alpha$:

$$P\{y \in \mathcal{C}(\boldsymbol{x})\} \geq 1 - \alpha \tag{1}$$

To construct the prediction sets, we split $\mathcal{I}$ into two disjoint subsets: a training set $\mathcal{I}_{\text{tr}}$ where we fit the classification model $f$, and a calibration set $\mathcal{I}_{\text{cal}} = \mathcal{I}/\mathcal{I}_{\text{tr}}$. Then, we get a fitted model $f$ on the $\mathcal{I}_{\text{tr}}$ and introduce a *non-conformity score* with a pre-defined score function $\mathcal{S}(\boldsymbol{x}, y) \in \mathbb{R}$ for each calibration data. We define a threshold $\tau$ to be the $1 - \alpha$ quantile of the non-conformity scores $\{s_i := \mathcal{S}(\boldsymbol{x}_i, y_i)\}_{i \in \mathcal{I}_{\text{cal}}}$. Consequently, we generate the prediction set for a test sample $\boldsymbol{x}_{\text{test}}$, containing labels whose score is under the threshold:

$$\mathcal{C}(\boldsymbol{x}_{\text{test}}) = \{y \in \mathcal{Y} : \mathcal{S}(\boldsymbol{x}_{\text{test}}, y) \leq \tau\} \tag{2}$$

As stated before, the prediction sets are provable to satisfy the coverage property in Eq.1. This coverage guarantee ([15]) is given by the following theorem:

**Theorem 2.1.** *Suppose* $(\boldsymbol{x}_i, y_i)_{i=1}^n$ *and* $(\boldsymbol{x}_{\text{test}}, y_{\text{test}})$ *are exchangeable data samples. Define* $\tau$ *as*

$$\tau = \inf\left\{ s : \frac{|\{i : \mathcal{S}(\boldsymbol{x}_i, y_i) \leq s\}|}{n} \geq \frac{\lceil (n+1)(1-\alpha) \rceil}{n} \right\}$$

*and the resulting prediction sets as*

$$\mathcal{C}(\boldsymbol{x}) = \{y : \mathcal{S}(\boldsymbol{x}, y) \leq \tau\}$$

*Then,*

$$P\{Y_{\text{test}} \in \mathcal{C}(\boldsymbol{x}_{\text{test}})\} \geq 1 - \alpha$$

Now, we review a commonly used non-conformity score function: **Adaptive Prediction Set (APS, [12])**. In the APS method, the non-conformity scores are calculated by accumulating sorted softmax probabilities

$$\mathcal{S}(\boldsymbol{x}, y) = \pi_{(1)}(\boldsymbol{x}) + \cdots + u \cdot \pi_{o(y, \pi(\boldsymbol{x}))}(\boldsymbol{x}) \tag{3}$$

where $\pi_{(1)}(\boldsymbol{x}), \pi_{(2)}(\boldsymbol{x}), \cdots, \pi_{(k)}(\boldsymbol{x})$ are the sorted softmax probabilities from greatest to smallest, $o(y, \pi(\boldsymbol{x}))$ represents the order of $\pi_y(\boldsymbol{x})$ and $u$ is a uniform random variable in $[0, 1]$ to break ties.

## 2.3 Evaluation metrics

The primary metrics used for evaluation are *coverage* and *inefficiency*. Coverage measures the percentage of samples whose prediction sets contain true labels:

$$\text{Coverage} = \frac{1}{|\mathcal{I}_{\text{test}}|} \sum_{i \in \mathcal{I}_{\text{test}}} \mathbf{1}\{y_i \in \mathcal{C}(\boldsymbol{x}_i)\} \tag{4}$$

where $\mathbf{1}\{E\}$ is the indicator function. Inefficiency or $N$ *criterion* is a measurement of the average size of prediction sets:

$$\text{Inefficiency} = \frac{1}{|\mathcal{I}_{\text{test}}|} \sum_{i \in \mathcal{I}_{\text{test}}} |\mathcal{C}(\boldsymbol{x}_i)| \tag{5}$$

From the point that the prediction set should both provide adequate coverage and be informative, sets of small size are preferred since they convey more detailed information and are more useful in practice.

# 3 Experiment

With the use of the vision-language model in real-world applications [10, 16], the assessment of VLM's uncertainty becomes increasingly important. Employed with conformal prediction, VLM can produce prediction sets under user-specified error, thus enhancing its security and reliability. However, this employment is not a trivial task since it's uncertain whether prediction sets can maintain the coverage rate in the zero-shot setting with different vision backbones. Moreover, in the task of medical image identification where data samples are limited and thus the calibration set is small, how will prediction sets perform remains unknown. Therefore, we focus on the following three problems: (1) **will prediction sets violate coverage rate under the zero-shot setting?** (2) **how do different vision backbones affect prediction sets?** (1) **how does the size of the calibration dataset influence prediction sets?**

## 3.1 Experiment Setup

The performance of conformal prediction is evaluated on CIFAR-10 [6] with four different vision backbones: ResNet50, ResNet101 [5], ViT-B-16 and ViT-B-32 [4]. To answer the first two questions, We divide the corresponding test dataset equally into a calibration set containing 5000 images and a test set containing 5000 images. For the second question, we evaluate the coverage and inefficiency of prediction sets with varying calibration set sizes in (50 100 500 1000 1500 2000). For each experiment, we repeat it 10 times and report the average value. The error rate $\alpha$ is set to be 0.05.

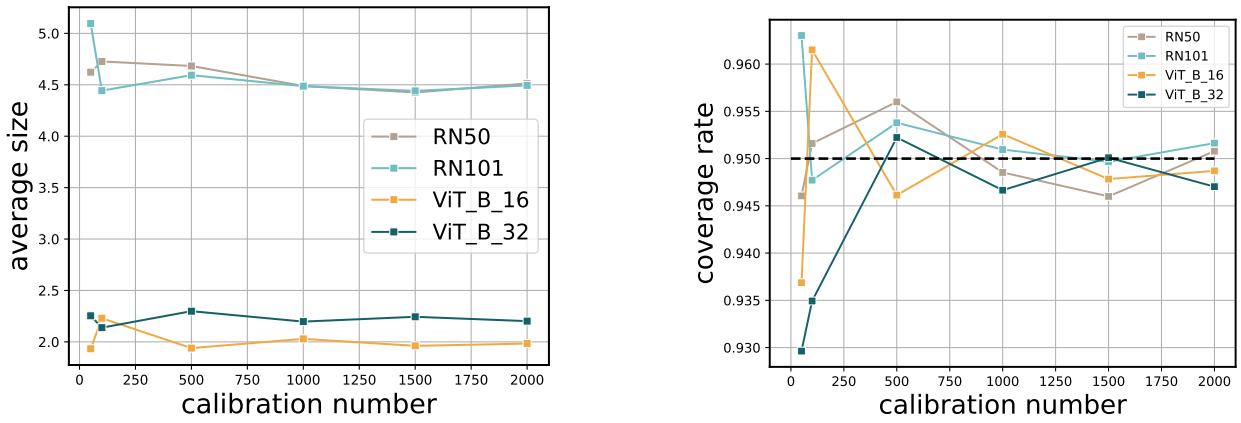
## 3.2 Empirical Results

Table 1: Comparison results of the coverage and average size of prediction sets with different vision backbones.

| Model | size | coverage | acc |
|---|---|---|---|
| rn50 | 4.48 | 0.95 | 0.57 |
| rn101 | 4.47 | 0.95 | 0.58 |
| vit16 | 1.99 | 0.95 | 0.87 |
| vit32 | 2.21 | 0.95 | 0.86 |

**Prediction sets maintain coverage rate under zero-shot setting with four vision backbones.** In Table 1, we compare the coverage rate of prediction sets with four different vision backbones. The results show that all vision backbones generate prediction sets that satisfy the required error rate 0.05. We give this phenomenon an intuitive explanation: the zero-shot method does not affect the exchangeability of data samples, and thus the Theorem 2.1 guarantees the coverage rate of prediction sets.

**The vision backbone that has high accuracy generates smaller prediction sets.** We also report the average size and accuracy of four vision backbones in Table 1, demonstrating that high accuracy often correlates with smaller prediction sets. This is because high accuracy indicated less uncertainty, thereby leading to prediction sets with small sizes.




(a) average size

(b) coverage

Figure 1: Comparison results of different calibration set sizes.

**Small calibration number may cause models to violate coverage rate.** In Figure 1, we compare how prediction sets perform with different calibration dataset sizes. The results show that conformal prediction tends to violate coverage rate with a small calibration number (either exhibits high coverage rate or low coverage rate). For example, on RN101, the coverage rate is over 0.96 (much higher than the required coverage 0.95), which is caused by large prediction sets. Moreover, on ViT_B_32, even though the average size of prediction sets approximates the result in Table 1, their coverage rate is 0.93.

# References


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

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

[3] Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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

[6] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[7] Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*, 2023.

[8] Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[9] Jing Lei and Larry Wasserman. Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96, 2014.

[10] Ziyuan Qin, Huahui Yi, Qicheng Lao, and Kang Li. Medical image understanding with pretrained vision language models: A comprehensive study. *arXiv preprint arXiv:2209.15517*, 2022.

[11] Allen Z Ren, Anushri Dixit, Alexandra Bodrova, Sumeet Singh, Stephen Tu, Noah Brown, Peng Xu, Leila Takayama, Fei Xia, Jake Varley, et al. Robots that ask for help: Uncertainty alignment for large language model planners. *arXiv preprint arXiv:2307.01928*, 2023.

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

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

[14] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

[15] Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. 1999.

[16] Xingcheng Zhou, Mingyu Liu, Bare Luka Zagar, Ekim Yurtsever, and Alois C Knoll. Vision language models in autonomous driving and intelligent transportation systems. *arXiv preprint arXiv:2310.14414*, 2023.