A Framework for Efficient Model Evaluation through Stratification, Sampling, and Estimation

Riccardo Fogliato[†] Pratik Patil[‡] Mathew Monfort[†] Pietro Perona[†§]

Abstract

Model performance evaluation is a critical and expensive task in machine learning and computer vision. Without clear guidelines, practitioners often estimate model accuracy using a one-time completely random selection of the data. However, by employing tailored sampling and estimation strategies, one can obtain more precise estimates and reduce annotation costs. In this paper, we propose a statistical framework for model evaluation that includes stratification, sampling, and estimation components. We examine the statistical properties of each component and evaluate their efficiency (precision). One key result of our work is that stratification via k-means clustering based on accurate predictions of model performance yields efficient estimators. Our experiments on computer vision datasets show that this method consistently provides more precise accuracy estimates than the traditional simple random sampling, even with substantial efficiency gains of 10x. We also find that model-assisted estimators, which leverage predictions of model accuracy on the unlabeled portion of the dataset, are generally more efficient than the traditional estimates based solely on the labeled data.

1 Introduction

Measuring the accuracy of computer vision (CV) algorithms is necessary to compare different approaches and to deploy systems responsibly. Yet, data labeling is expensive. While machine learning techniques are increasingly able to digest large *training* sets that are sparsely and noisily annotated, *test* sets require a greater level of care in their construction. First, the tolerance for annotation quality is much stricter, as annotation errors will lead to an incorrect estimation of model accuracy. Second, data must be collected and annotated at a scale such that the confidence intervals around the error rates are sufficiently narrow (compared to the error rates) to make meaningful comparisons between models and error rates have been plummeting. Lastly, for many applications, evaluating a single model can involve multiple test sets designed to assess performance in different domains, metrics, and scenarios. This is necessary in testing, for example, cross-modal models such as CLIP [77]. Practitioners facing the cost of putting together test sets will ask a simple question: *How can one minimize the number of annotated test samples that are required to precisely estimate the predictive accuracy of a model?*

Efficient estimation of model accuracy can be achieved by co-designing sampling strategies (for selecting which data points to label) and statistical estimation strategies (for calculating model performance). One may craft sampling strategies that maximize the (statistical) efficiency of a given method for estimating model accuracy, that is, minimize its error given a fixed number of annotated samples [47, 55]. Unlike simple random sampling, which picks any example from the dataset with

[†]Amazon Web Services; corresponding author email: fogliato@amazon.com
[‡]University of California Berkeley
[§]California Institute of Technology

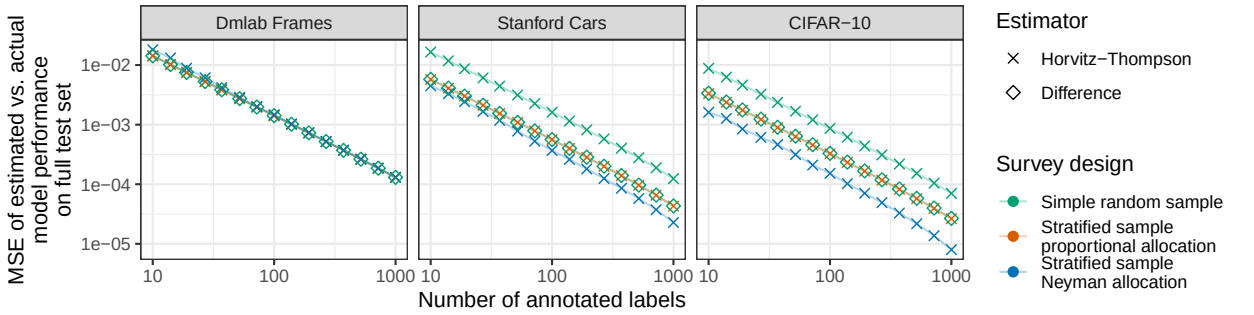

Figure 1: **Mean squared errors (MSEs) of estimators across sampling designs.** Estimates of zero-shot accuracy of ViT-B/32 in classification tasks on three datasets as a function of the amount of labeled data. Stratified sampling can dramatically reduce the number of annotations needed to accurately estimate the model accuracy compared to the naive average (`HT`) under simple random sampling. Neyman allocation can sometimes further improve precision compared to proportional allocation. (From left to right) No savings on the Dmlab Frames dataset, about 5x savings on the Stanford Cars, and about 10x savings on CIFAR-10. Note that the efficiency (precision) gains vary considerably between datasets (analysis and discussion in Section 5). In the absence of stratified sampling with k-means on model predictions, the difference estimator can also greatly help.

equal probability, efficient strategies will select the most informative instances to annotate when constructing a test set. One may also look for efficient estimators. Unlike design-based approaches, which base the statistical inference solely on the labeled sample, model-assisted estimators leverage the predicted labels on the remaining data to increase the precision of the estimates [8, 100, 82]. However, CV researchers continue to rely on simple random sampling and design-based inference. Why?

We believe that there are two reasons why efficient sampling strategies and estimators have not yet been adopted. First, although the literature offers many different statistical techniques, CV practitioners do not have guidance towards a "backpocket method" that they can trust out-of-the-box. Second, there is no comprehensive study that compares sampling and estimation strategies on CV data. Thus, it is not clear whether the additional complexity of using such sampling techniques will pay off in terms of lower costs.

We address both issues here. We aim to give a readable and systematic account of methods from the statistics literature, test them on a large palette of CV models and datasets, and make a final recommendation for a simple and efficient method that the community can readily adopt. We take a practical point of view and choose to focus on one-shot selection techniques, rather than sequential sampling. This is because the job of annotating data is typically contracted out and carried out all at once, and thus the process of data sampling has to take place entirely before data annotation.

More specifically, our work outlines a framework consisting of stratification, sampling design, and estimation components that practitioners can utilize when evaluating model performance. We review simple and stratified random sampling strategies with proportional and (optimal) Neyman allocation, as well as the Horvitz-Thompson and model-assisted difference estimators [8]. Building on the survey sampling literature, we describe how to stratify the sample and design sampling strategies tailored to maximize the efficiency of the target estimator. We show that one should leverage accurate predictions of model performance (e.g., the predicted classification error of a CV classifier) in the stratification procedure or via the difference estimator to increase the precision of the estimates and reduce the number of samples needed for testing. We experimentally show how to apply the

framework to benchmark models on CV classification tasks.

Figure 1 shows the main takeaways from our work. The model-assisted difference estimator and stratified sampling strategies (both proportional and Neyman) can significantly improve the precision of CV classifier accuracy estimates compared to naive averaging (Horvitz-Thompson) on a random data subset, e.g., achieving a 10x gain in precision on CIFAR-10 [57]. While the improvements may vary (e.g., modest gains on Stanford Cars [56]) or be less pronounced in some cases (e.g., DMlab Frames [104]), stratified sampling with proportional allocation consistently offers a reliable and often superior performance. Can one predict when clever methods will yield more bang for the buck? Yes, in Section 5, we explore the question of when and why these methods provide the most benefit to gain insights to apply them effectively.

Contributions and outline. A summary of our contributions and paper outline is as follows:

- In Section 3, we prescribe a statistical framework for model evaluation consisting of stratification, sampling, and estimation components. (Algorithm 1).
- In Section 4, we discuss the design of stratification, sampling, and estimator. In particular, we show that maximizing the efficiency of the Horvitz-Thompson estimator under proportional allocation is equivalent to optimizing a k-means criterion (Proposition 2 and Corollary 3).
- In Section 5, we explore the behavior of different options using a wide range of experiments on CV datasets. We find that carefully designed stratification strategies as well as model-assisted estimators always yield more precise estimates of model performance compared to naive estimation under simple random sampling (Figure 2). Calibration and accurate prediction of the loss are key to obtaining highly efficient estimators (Figure 3).

2 Related Work

The idea of using clever sampling and estimators to obtain more precise estimates of a target of interest on a dataset has been extensively studied in the fields of survey sampling and machine learning. We review the relevant literature in these areas below.

2.1 Related Work in Survey Sampling

The question of efficient or precise evaluation is essentially analogous to problems encountered in survey sampling [30, 83, 65, 21, 35, 48] Survey sampling has two main inference paradigms. The first, design-based inference, views the dataset as static and assumes randomness only in sample selection. A well-known estimator in this framework, which we focus on, is the Horvitz-Thompson estimator [43], which averages the labeled samples reweighed by their propensity to be sampled. The second, model-based inference, assumes that the data are drawn from a superpopulation and uses statistical models for inference, leading to more precise estimates when the model is well-specified and less precise estimates when the model is misspecified. The model-assisted approach combines the strengths of both by integrating modeling into the design-based framework. This approach yields (nearly) unbiased estimates as in the design-based paradigm but that are more precise when the model is correct. We focus on a popular model-assisted estimator, the difference estimator [66, 100, 83, 82]. Our findings align with existing survey sampling literature [8, 7, 69], demonstrating that model-assisted estimators can significantly improve the precision of model performance estimates when predictions on the unlabeled sample are accurate.

Another crucial element in survey sampling is the design of the sample collection itself, which should aim to maximize the efficiency of the target estimator [37, 12, 11, 45, 19]. There is a wide range of sampling designs (i.e., probability distributions over all possible samples), each designed to meet specific needs and contexts, together with the corresponding estimators [94, 9]. In this paper, we focus on simple random sampling with and without stratification because of its easily understandable advantages and trade-offs. We bypass more complex strategies such as unequal probability sampling, which can be carried out along with stratification, as they offer minimal additional benefits compared to stratified sampling with Neyman allocation when the number of strata is large [73].

2.2 Related Work in Machine Learning

Efficient data sampling and estimation techniques have also been extensively discussed in the machine learning literature, particularly in the following settings.

Model performance estimation with fewer labels. Multiple works have employed design-based estimators and considered the (active) setting where the labels are sampled iteratively [63, 47]. The devised sampling designs generally rely on stratification or unequal probability sampling, using predictions of model accuracy generated by the model itself [84, 85, 76] or by a surrogate model [55, 54]. While our work shares many similarities with this line of research, we specifically focus on scenarios where labels are selected simultaneously, (mathematically and empirically) compare findings from different classes of estimators, offering practical advice on the best way to stratify. In addition, while these works focus on a few selected datasets, we compare the methods through a comprehensive array of experiments (see Section 5).

Model performance estimation on unlabeled or partially labeled data. Our paper is related to efforts on the estimation of model performance on unlabeled data [10, 24, 99, 71, 13, 103, 15]. These works focus on the prediction of classification accuracy on out-of-distribution data, leveraging indicators of distribution shift between training and test data such as Fréchet distance [25], discrepancies in model confidence scores between validation and test data [36, 33], and disagreement between the predictions made by multiple similar models [17, 49, 4]. A key takeaway from these works is that accurate estimation is a byproduct of proper model calibration [96], which is itself an area of active research [51, 80, 39]. Some studies also address this challenge using a mix of unlabeled and labeled data, applying parametric models to predictions and existing labels [98, 70]. Notably, recent research has explored "prediction-powered" inference, a class of estimators that uses model predictions on the unlabeled data in the estimation process [1, 2, 105, 106]. In the case of mean estimation, this coincides with the model-assisted difference estimator from survey sampling. This line of work focuses on simple random sampling and Poisson sampling designs. We contribute to this literature by comparing the performance of the difference estimator across stratified sampling methods. Our results show that, when stratified designs are used, the difference and Horvitz-Thompson estimators perform similarly.

Active learning. Our paper is also related to the literature on pool-based active learning, where the goal is to minimize the number of labels that are needed to ensure that the model achieves a given predictive accuracy [89, 22, 61, 16, 90, 32, 28]. This is done by iterating between sampling and retraining. Traditionally, sampling designs in this area have focused on the predictive uncertainty of the model [46], selecting instances one at a time [62, 86]. More recent work has explored other approaches [31, 88, 79] and batch sampling strategies [52, 3]. Sampling strategies for model training and evaluation share many similarities. However, while (optimal) sampling designs tailored towards

evaluation prioritize the sampling of data where model performance is most uncertain, active learning sampling approaches favor the sampling of observations that are anticipated to boost model performance.

3 Framework Overview

We provide a formal description of the problem setup and of our framework in Section 3.1 and Section 3.2 respectively. To ground our discussion, we use a classification task as a recurring example, although our framework also applies to regression tasks.

3.1 Formal Setup

Consider a dataset $\mathcal{D}$ consisting of N instances $\{(X_i, Y_i)\colon i = 1, \ldots, n\}$ drawn independently from distribution P. (Think of each instance as an image $X_i \in \mathcal{X}$ and its corresponding ground truth label $Y_i \in \mathcal{Y}$.) We have access to a predictive model f that outputs estimates $f_y(X_i)$ of the likelihood that label $y \in \mathcal{Y}$ is present in the i-th image X_i for all $y \in \mathcal{Y}$. The predicted label with the highest score is $\widehat{Y}_i = \arg\max_{y \in \mathcal{Y}} f_y(X_i)$. Let (X, Y) be a draw from P and let Z be the predictive error of our model f on (X, Y). Our target of interest is a predictive performance metric θ of the model f, defined as $\theta = \mathbb{E}_P[Z]$. For example, taking $Z = \mathbb{1}(Y = \widehat{Y})$ yields the usual classification accuracy, $Z = (1 - f_Y(X))^2$ the squared error, and $Z = -\log f_Y(X)$ the cross-entropy.

In principle, we could estimate θ using $\mathcal{D}$ by $\widehat{\theta}_{\mathcal{D}} = N^{-1} \sum_{i \in \mathcal{D}} Z_i$. However, while we have access to X and to the outputs of f for all $1 \leq i \leq N$, Y is not readily available. Our budget only allows us to obtain Y for a subset of the instances $\mathcal{S} \subset \mathcal{D}$ of size $n \ll N$. We will randomly select these instances according to a sampling design π, which is a probability distribution over all subsets of size n in $\mathcal{D}$. We denote by $\pi_i > 0$ the likelihood that the i-th instance is included in $\mathcal{S}$. Using the available data, we then obtain an estimate $\widehat{\theta}$ of $\widehat{\theta}_{\mathcal{D}}$.

We measure the *efficiency* of the estimator $\widehat{\theta}$ of θ in terms of its mean squared error $\mathrm{MSE}(\widehat{\theta}, \theta) = \mathbb{E}_P[\mathbb{E}_\pi[(\widehat{\theta} - \theta)^2]]$. The bias-variance decomposition yields $\mathrm{MSE}(\widehat{\theta}, \theta) \approx (\mathbb{E}_P[\mathbb{E}_\pi[\widehat{\theta}]] - \widehat{\theta}_{\mathcal{D}})^2 + \mathbb{E}_P[\mathrm{Var}_\pi(\widehat{\theta})]$ because $\mathrm{Var}_P(\mathbb{E}_\pi[\widehat{\theta}])$ is small when $n \ll N$. Thus, the MSE will be driven by the bias and variance over the sampling design. Since we will only look at design-unbiased estimators, the MSE will correspond to the variance. We define the relative efficiency of estimator $\widehat{\theta}^{(1)}$ relative to $\widehat{\theta}^{(2)}$ under a sampling design π as the inverse of the ratio of their MSEs, that $\mathrm{MSE}_\pi(\widehat{\theta}^{(2)}, \widehat{\theta}_{\mathcal{D}}) / \mathrm{MSE}_\pi(\widehat{\theta}^{(1)}, \widehat{\theta}_{\mathcal{D}})$. We say that estimator $\widehat{\theta}^{(1)}$ is more efficient than $\widehat{\theta}^{(2)}$ when the relative efficiency is greater than one.

3.2 Framework Overview

Algorithm 1 outlines a framework for estimating the performance of a predictive model from a dataset $\mathcal{D}$ when only a subset $\mathcal{S}$ of instances has been labeled. The framework consists of an optional step for predicting model performance, a stratification or clustering procedure, a sampling design or strategy, and an estimator. Next, we discuss the choices for each of these components.

Prediction (of Z). The first step involves building a proxy $\widehat{Z}$ of Z that is *independent* of the observed labels. For example, when $Z = \mathbb{1}(Y = \widehat{Y})$ represents the accuracy of the classifier, we could take $\widehat{Z}_i := \mathbb{E}[Z_i \,|\, X_i] = \mathbb{P}(Y_i = \widehat{Y}_i)$, which can be estimated via:

- *Model predictions* $f(X)$: Use $\widehat{Z}_i = f_{\widehat{Y}_i}(X_i)$ (likelihood of the model's top class) as a proxy.

Algorithm 1 A Framework for Efficient Model Evaluation (see Section 3.2)

Input: Test dataset $\mathcal{D}$ of size N with predictions of f, annotation budget $n \ll N$.

1: **Predict**: Construct a proxy $\widehat{Z}$ of $\mathbb{E}_P[Z \mid X]$ and add predictions $\{\widehat{Z}_i\}_{i\in\mathcal{D}}$ to $\mathcal{D}$.
2: **Stratify**: Partition the dataset into H strata (or clusters) $\{\mathcal{D}_h\}_{h=1}^H$ using $\widehat{Z}$ or X.
3: **Sample**: Select $\mathcal{S}$ ($|\mathcal{S}| = n$) from $\mathcal{D}$ based on the chosen design.
4: **Annotate**: Obtain labels $\{Y_i\}_{i\in\mathcal{S}}$, compute performance $\{Z_i\}_{i\in\mathcal{S}}$.
5: **Estimate**: Compute estimate $\widehat{\theta}$ of model performance $\theta = \mathbb{E}_P[Z]$.

Output: Estimate $\widehat{\theta}$.

- *Auxiliary predictions* $f^*(X)$: Use $\widehat{Z}_i = f^*_{\widehat{Y}_i}(X_i)$, the prediction of an auxiliary model f^* that, similarly to f, estimates the probability distribution of Y.

Stratification. Stratification involves partitioning the population $\mathcal{D}$ into $H > 0$ strata, $\{\mathcal{D}_h\}_{h=1}^H$ with $|\mathcal{D}_h| = N_h$. We can use standard clustering algorithms to form the strata based on:

- *Proxy* $\widehat{Z}$: Construct strata using the estimates $\{\widehat{Z}_i\}_{i\in\mathcal{D}}$.
- *Features* X: Cluster the images $\{X_i\}_{i\in\mathcal{D}}$, e.g., by using their feature representations obtained from an encoder.

Sampling. Two popular classes of sampling designs with fixed size and without replacement are:

- *Simple random sampling* (`SRS`): Randomly sample n instances from $\mathcal{D}$ with equal probability.
- *Stratified simple random sampling* (`SSRS`): Allocate budget n_h to each stratum $1 \le h \le H$ such that $\sum_{h=1}^H n_h = n$ and conduct `SRS` within each stratum, obtaining $\mathcal{S}_h$. `SSRS` designs differ in how the budget n is allocated to strata. We analyze two allocation strategies in Section 4. Throughout our discussion, we will assume that strata sizes are large: $1/N_h \approx 0\ \forall h \in [H]$.

Estimation. We consider two instances of (unbiased) design-based and model-assisted estimators, chosen for their well-established statistical properties.[1]

- *Horvitz-Thompson estimator* (`HT`) [43]: This design-based estimator is defined as:

$$\widehat{\theta}_{\texttt{HT}} = \frac{1}{N}\sum_{i\in\mathcal{S}}\frac{Z_i}{\pi_i}. \tag{1}$$

  `HT` is design-unbiased, that is, $\mathbb{E}_\pi[\widehat{\theta}_{\texttt{HT}}] = \widehat{\theta}_{\mathcal{D}}$ for $\pi \in \{\texttt{SRS}, \texttt{SSRS}\}$. It follows that $\mathrm{MSE}(\widehat{\theta}, \theta) \approx \mathbb{E}_P[\mathrm{MSE}(\widehat{\theta}, \widehat{\theta}_{\mathcal{D}})] = \mathbb{E}_P[\mathrm{Var}_\pi(\widehat{\theta})]$ and the design-based variance is the sole source of error.

- *Difference estimator* (`DF`) [83, 8]: This model-assisted estimator is defined as:

$$\widehat{\theta}_{\mathrm{DF}} = \frac{1}{N}\sum_{i\in\mathcal{D}}\widehat{Z}_i + \frac{1}{N}\sum_{i\in\mathcal{S}}\frac{Z_i - \widehat{Z}_i}{\pi_i}, \tag{2}$$

[1] In preliminary experiments we assessed other model-assisted estimators in the class of "generalized" regression estimators [83, 100, 8] and found results comparable to `DF`.

where $\widehat{Z}_i$ is an estimate of Z_i. The first term, $\sum_{i\in\mathcal{D}} \widehat{Z}_i$, is independent of the sampling strategy. The second term corrects the bias of the first term, as $\sum_{i\in\mathcal{S}} \mathbb{E}_\pi[(Z_i - \widehat{Z}_i)/\pi_i] = \sum_{i\in\mathcal{D}}(Z_i - \widehat{Z}_i)$. This makes the `DF` estimator also unbiased under the sampling design.

These estimators offer complementary strengths: `HT` offers simplicity and unbiasedness, while `DF` provides potential variance reduction by incorporating model predictions. Our framework leverages these properties to improve the efficiency of model performance evaluation in computer vision tasks. In the next section, we will discuss the optimal design of these stratification, sampling, and estimation components.

4 Design of Framework Components

To evaluate the effectiveness of the components in determining the estimator's variance or efficiency, we review the optimality of each. In Section 4.1 we analyze the efficiency of the estimators under the sampling designs. In Section 4.2 we discuss how the choice of the proxy $\widehat{Z}$ for Z can improve the efficiency of stratified sampling procedures. Lastly, in Section 4.3 we discuss the choice of the proxy in terms of the efficiency of the `DF` estimator.

4.1 Choosing the Sampling Design

Under `SRS`, $\pi_i = n/N$ for all $1 \le i \le N$ and the `HT` estimator is simply the traditional empirical average $n^{-1}\sum_{i\in\mathcal{S}} Z_i$. Its MSE under the sample design is given by:

$$\mathrm{MSE}_{\texttt{SRS}}(\widehat{\theta}_{\texttt{HT}}, \widehat{\theta}_{\mathcal{D}}) = \frac{1-f}{n} S_Z^2, \tag{3}$$

where $f = n/N$ represents sampling fraction and $S_Z^2 = (N-1)^{-1}\sum_{i\in\mathcal{D}}(Z_i - \widehat{\theta}_{\mathcal{D}})^2$ is the variance of Z in the finite population. From standard arguments in sampling statistics, under the setup of Section 3, it can be shown that

$$\mathrm{MSE}_{\texttt{SRS}}(\widehat{\theta}_{\texttt{HT}}, \widehat{\theta}_{\mathcal{D}})^{-1/2}(\widehat{\theta}_{\texttt{HT}} - \widehat{\theta}_{\mathcal{D}}) \xrightarrow{d} \mathcal{N}(0,1)$$

as $n, N \to \infty$ and $N - n \to 0$ (see, e.g., Corollary 1.3.2.1 in [30]). Estimation of the uncertainty around $\widehat{\theta}_{\texttt{HT}}$ can be performed using a plug-in estimator of the variance S_Z^2 in (3). In particular, when $f \approx 0$ and $1/n \approx 0$, we recover the common MSE or variance estimator $\mathrm{MSE}_{\texttt{SRS}}(\widehat{\theta}_{\texttt{HT}}, \widehat{\theta}_{\mathcal{D}}) \approx \sum_{i\in\mathcal{S}}(Z_i - \widehat{\theta}_{\texttt{HT}})^2/n^2$.

One standard `SSRS` approach to budget splitting is proportional allocation, which assigns the budget proportionally to the size of the stratum in the finite population. For all $1 \le h \le H$, we assign $n_h \propto N_h$ and set $\pi_i = n_h/N_h$ for all $i \in \mathcal{D}_h$. Under this allocation, the `HT` estimator is $\widehat{\theta}_{\texttt{HT}} = N^{-1}\sum_{h=1}^{H}(N_h/n_h)\sum_{i\in\mathcal{S}_h} Z_i$ and its MSE is given by:

$$\mathrm{MSE}_{\texttt{SSRS},p}(\widehat{\theta}_{\texttt{HT}}, \widehat{\theta}_{\mathcal{D}}) = \frac{1-f}{n}\sum_{h=1}^{H}\frac{N_h}{N} S_{Z_h}^2, \tag{4}$$

where $S_{Z_h}^2$ is the variance of Z in the h-th stratum [94]. Analogous to `SRS`, asymptotic guarantees for `HT` under `SSRS` can also be obtained (see Theorem 1.3.2 in [30]).

One can also seek a budget allocation that minimizes the error of `HT`, which is $\mathrm{MSE}(\widehat{\theta}_{\texttt{HT}}, \widehat{\theta}_{\mathcal{D}})$. This strategy is known as Neyman or optimal allocation [73] and in the case of the `HT` estimator it assigns $n_h \propto N_h\sqrt{S_{Z_h}^2}$ [94, 21]. This means that more samples will be assigned to larger and more variable

strata compared to proportional sampling. The HT estimator remains the same as under proportional allocation but its MSE now becomes:

$$\mathrm{MSE}_{\texttt{SSRS},o}(\widehat{\theta}_{\texttt{HT}}, \widehat{\theta}_{\mathcal{D}}) = \frac{1}{n}\left(\sum_{h=1}^{N} \frac{N_h}{N} S_{Z_h}\right)^2 - \frac{1}{N}\sum_{h=1}^{N} \frac{N_h}{N} S^2_{Z_h}. \tag{5}$$

Since $\widehat{Z}$ does not depend on the labels in $\mathcal{S}$, the MSEs of the DF estimator under SRS and SSRS are obtained by replacing Z with $(Z - \widehat{Z})$ in the formulas above, including for Neyman allocation [12].

By comparing (3), (4), and (5), we can derive the following well-known result, which identifies the sampling designs that yield the most precise estimates of $\widehat{\theta}_{\mathcal{D}}$ [21, 94, 65, 30].

Proposition 1. *Under the setup of Section 3,*

$$\mathrm{MSE}_{\texttt{SSRS},o}(\widehat{\theta}_{\texttt{HT}}, \widehat{\theta}_{\mathcal{D}}) \leq \mathrm{MSE}_{\texttt{SSRS},p}(\widehat{\theta}_{\texttt{HT}}, \widehat{\theta}_{\mathcal{D}}) \leq \mathrm{MSE}_{\texttt{SRS}}(\widehat{\theta}_{\texttt{HT}}, \widehat{\theta}_{\mathcal{D}}). \tag{6}$$

Similar inequalities also hold for the DF estimator. This result establishes that SSRS with proportional allocation consistently yields estimates with equal or lower MSE compared to SRS for the HT and DF estimators. The reduction in MSE depends on the homogeneity of the strata: When model performances Z within each stratum are mostly equal, gains in efficiency of SSRS compared to SRS are largest. When we know the standard deviation S_{Z_h} and this term varies substantially across strata, Neyman allocation can provide even more precise estimates than proportional sampling. However, when our estimates of S_{Z_h} are incorrect, Neyman allocation may lead to less precise even compared to SRS. The empirical results presented in Section 5 align with these conclusions.

4.2 Designing the Strata

We turn to the construction of the strata. We can rewrite the MSE of the HT estimator under SSRS in (4) with proportional allocation as:

$$\mathrm{MSE}_{\texttt{SSRS},p}(\widehat{\theta}_{\texttt{HT}}, \widehat{\theta}_{\mathcal{D}}) \approx \frac{1-f}{Nn}\left\{\sum_{h=1}^{H}\sum_{i\in\mathcal{D}_h}\left[(\widehat{Z}_i - \widehat{\overline{Z}}_{\mathcal{D}_h})^2 + (\widehat{\theta}^2_{\mathcal{D}_h} - \widehat{\overline{Z}}^2_{\mathcal{D}_h})\right] + \sum_{i\in\mathcal{D}}\left[(Z_i - \widehat{Z}_i)^2 + 2\widehat{Z}_i(Z_i - \widehat{Z}_i)\right]\right\}, \tag{7}$$

where $\widehat{\overline{Z}}_{\mathcal{D}_h} = N_h^{-1}\sum_{i\in\mathcal{D}_h}\widehat{Z}_i$ and $\widehat{\theta}_{\mathcal{D}_h} = N_h^{-1}\sum_{i\in\mathcal{D}_h} Z_i$. The first term on the right-hand side of (7) represents the within-strata sum of squares of the predictions $\{\widehat{Z}_i\}_{i\in\mathcal{D}}$. When $\widehat{Z}_i \approx Z_i$, the second term becomes negligible. Since the remaining terms do not depend on the stratification, the strata construction affects the MSE only through the first term. This intuition is formalized in the following result.

Proposition 2. *Assume that $\widehat{Z}_i = \mathbb{E}_P[Z_i \,|\, X_i]$ for all $i \in \mathcal{D}$. Then the partition $\{\mathcal{D}_h\}_{h=1}^H$ of $\mathcal{D}$ that minimizes $\sum_{h=1}^H (N_h/N) S^2_{\widehat{Z}h}$ also minimizes the error $\mathbb{E}_P[\mathrm{MSE}_{\texttt{SSRS},p}(\widehat{\theta}_{\texttt{HT}}, \widehat{\theta}_{\mathcal{D}}) \,|\, X]$ where $X = \{X_i\}_{i\in\mathcal{D}}$.*

The result follows from standard decompositions for proper scoring rules [58] and implies that minimizing the weighted within-strata sum of squares for $\widehat{Z}$ will also minimize the MSE of the HT estimator under proportional stratification. In other words, this means that when a good predictor of the model performance based on X is available, using its predictions *alone* (as compared to clustering on X) will be sufficient to maximize the efficiency of the HT estimator. This also provides practical guidance on which criterion to optimize, as summarized in the following corollary.

Corollary 3. *The partition* $\{\mathcal{D}_h\}_{h=1}^H$ *of* $\mathcal{D}$ *that minimizes* $\mathbb{E}_P[\mathrm{MSE}_{\texttt{SSRS},p}(\widehat{\theta}_{\texttt{HT}}, \widehat{\theta}_{\mathcal{D}}) \mid X]$ *is the same as that optimized by* k*-means clustering on* $\{\mathbb{E}_P[Z_i \mid X_i]\}_{i\in\mathcal{D}}$.

The corollary follows directly from Proposition 2. Thus, we can expect larger efficiency gains for the HT estimator under SSRS with proportional allocation when strata are formed by solving the k-means clustering criterion on $\mathbb{E}_P[Z \mid X]$. In practice, this expectation is unknown and we have to rely on its proxy $\widehat{Z}$. A natural choice for the clustering algorithm is then to use the k-means algorithm itself on the proxy. Nonetheless, the experiments in Section 4 will show that even with estimated values, this approach still leads to better efficiency gains compared to stratifying based on the feature representations of X obtained using the same model.

4.3 Choosing the Estimator

Based on our discussion in Section 4.1, one might have guessed that the DF estimator will have lower MSE than the HT estimator when Z and $\widehat{Z}$ are positively associated. To formally characterize this intuition, consider SRS, under which the MSE of the DF estimator is:

$$\mathrm{MSE}_{\texttt{SRS}}(\widehat{\theta}_{\texttt{DF}}, \widehat{\theta}_{\mathcal{D}}) \approx \frac{1-f}{n}\left\{\frac{1}{N}\sum_{i\in\mathcal{D}}(Z_i - \widehat{Z}_i)^2 - (\widehat{\theta}_{\mathcal{D}} - \widehat{\overline{Z}})^2\right\}. \tag{8}$$

The first term on the right-hand side of (8) represents the MSE of $\widehat{Z}_i$ with respect to Z_i, while the second term represents a squared calibration error. It follows that choosing $\widehat{Z}_i = \mathbb{E}_P[Z_i \mid X_i]$ for all $1 \le i \le N$ minimizes the expected MSE of DF under SRS. This choice for $\widehat{Z}$ aligns with our recommendation from the stratification procedure and leads to the following result:

Proposition 4. *Assuming* $\widehat{Z}_i = \mathbb{E}_P[Z_i \mid X_i]$, *we have*

$$\frac{\mathbb{E}_P[\mathrm{MSE}_{\texttt{SRS}}(\widehat{\theta}_{\texttt{DF}}, \widehat{\theta}_{\mathcal{D}})]}{\mathbb{E}_P[\mathrm{MSE}_{\texttt{SRS}}(\widehat{\theta}_{\texttt{HT}}, \widehat{\theta}_{\mathcal{D}})]} = \frac{\mathbb{E}_P[\mathrm{Var}_P(Z \mid X)]}{\mathrm{Var}_P(Z)}. \tag{9}$$

Since $\mathrm{Var}_P(Z) = \mathbb{E}_P[\mathrm{Var}_P(Z \mid X)] + \mathrm{Var}_P(\mathbb{E}_P[Z \mid X])$ by the law of total variance, the ratio in Proposition 4 will always be less than 1. This means that the DF estimator will yield more precise estimates than HT as long as $\widehat{Z}$ is well specified. The efficiency gains of DF over HT under SRS will be highest when the auxiliary information X is predictive of Z, that is, when $\mathrm{Var}_P(\mathbb{E}_P[Z \mid X])$ is large.

Under SSRS with proportional allocation, we can similarly show that

$$\mathrm{MSE}_{\texttt{SSRS},p}(\widehat{\theta}_{\texttt{DF}}, \widehat{\theta}_{\mathcal{D}}) \;\approx\; \frac{1-f}{n}\left\{\frac{1}{N}\sum_{i\in\mathcal{D}}(Z_i - \widehat{Z}_i)^2 - \sum_{h=1}^{H}\frac{N_h}{N}(\widehat{\theta}_{\mathcal{D}_h} - \widehat{\overline{Z}}_{\mathcal{D}_h})^2\right\}.$$

Since the first term on the right-hand side will in general dominate when the proxy is calibrated, we should not expect significant efficiency gains of DF compared to HT under this sampling design when the strata are finegrained enough, i.e., $\widehat{Z}_i - \widehat{\overline{Z}}_{\mathcal{D}_h} \approx 0$ for all $i \in \mathcal{D}_h$. Thus, the uncertainty of the DF and HT estimates under SSRS will be close, i.e., $\mathrm{MSE}_{\texttt{SSRS},p}(\widehat{\theta}_{\texttt{DF}}, \widehat{\theta}_{\mathcal{D}}) \approx \mathrm{MSE}_{\texttt{SSRS},p}(\widehat{\theta}_{\texttt{HT}}, \widehat{\theta}_{\mathcal{D}})$.

5 Empirical Evaluation

5.1 Experimental Setup

To evaluate the methods, we consider the classification setup described in Section 3. Our goal is to compare the efficiency or precision of sampling designs and associated estimators of the predictive

performance of a model f, namely $\theta = \mathbb{E}_P[Z]$, by having access only to a limited number of labels (say $n = 100$) from our test dataset $\mathcal{D}$ of size $N \gg n$.

Tasks and models. Our main evaluation focuses on the zero-shot classification accuracy ($Z = \mathbb{1}(Y = \widehat{Y})$) of a CLIP model f with ViT-B/32 as the visual encoder, pretrained on the English subset of LAION-2B [44, 87, 77]. We evaluate its accuracy on the tasks included in the LAION CLIP-Benchmark [59]; the full list is provided in Appendix B. This benchmark covers a wide range of model performances and task diversities, making it a suitable testbed for comparing different estimation methods. To construct $\widehat{Z}$, we use the confidence scores from either CLIP ViT-B/32 or from the surrogate model f^* CLIP ViT-L/14. The latter model achieves higher classification accuracy than the former across most tasks in the benchmark, meaning that the proxy $\widehat{Z}$ is a better predictor of Z. Additionally, we calibrate the proxy $\widehat{Z}$ with respect to Z via isotonic regression on a randomly sampled half subset of $\mathcal{D}$; technically, one could conduct training and evaluation on the same dataset with cross-fitting. We carry out the estimation procedure on the remaining half of the data, $\mathcal{D}$. For stratification purposes, we obtain feature representations from the penultimate layer of f. We arbitrarily set the number of strata to 10 for all experiments; in the case of `SSRS` with proportional allocation, more strata would lead to more efficiency gains. Our code and package is available at github.com/amazon-science/ssepy.

Additional experiments. In Appendix B, we include additional experiments to evaluate the performance of our methods. These experiments cover: (Appendix B.2) the estimation of performance metrics other than classification accuracy; (Appendix B.3) results with predictions generated with linear probing and (Appendix B.4) with predictions by CLIP with ResNet and ConvNeXT backbones [38, 64]; (Appendix B.5) an analysis on two datasets from the WILDS out-of-distribution benchmark [53]. Some of the results of these experiments are also summarized in Section 6. In particular, we compare the efficiency of our methods on data that is out-of-distribution for the model and for the proxy $\widehat{Z}$ of Z.

5.2 Results

We study the efficiency of sampling design, stratification procedures, and estimators. We then analyze where the efficiency gains over `HT` under `SRS` arise.

Sampling design. Proposition 1 states that estimates obtained through `SSRS` with proportional allocation using the `HT` and `DF` estimators consistently achieve lower variance or MSE compared to those obtained via `SRS`. Figure 2 corroborates this analytical finding (see also the results in Table in Appendix B), showing that $\text{MSE}_{\texttt{SSRS}}(\widehat{\theta}_{\texttt{HT}}, \widehat{\theta}_{\mathcal{D}}) \leq \text{MSE}_{\texttt{SRS}}(\widehat{\theta}_{\texttt{HT}}, \widehat{\theta}_{\mathcal{D}})$ regardless of the features used for stratification. The gain varies across tasks and, when using surrogate model predictions, the relative efficiency ranges from about 10x on some tasks to no gain on others. While Neyman allocation is guaranteed to yield more precise estimates compared to these sampling designs when the allocation is based on S_{Z_h}, in practice we need to rely on its plug-in estimator $\widehat{S}_{Z_h} = [\widehat{\overline{Z}}_{\mathcal{D}_h}(1 - \widehat{\overline{Z}}_{\mathcal{D}_h})]^{1/2}$. This can introduce inaccuracies in the budget allocation. Indeed, we observe that Neyman allocation can perform even worse than `SRS` and `SSRS` with proportional allocation. However, when $\widehat{Z}$ is derived from the predictions of the surrogate model f^* and is further calibrated, then Neyman allocation consistently matches or exceeds the performance under `SRS`. On certain tasks, the MSE of `HT` is more than 10x lower compared to under `SRS`.

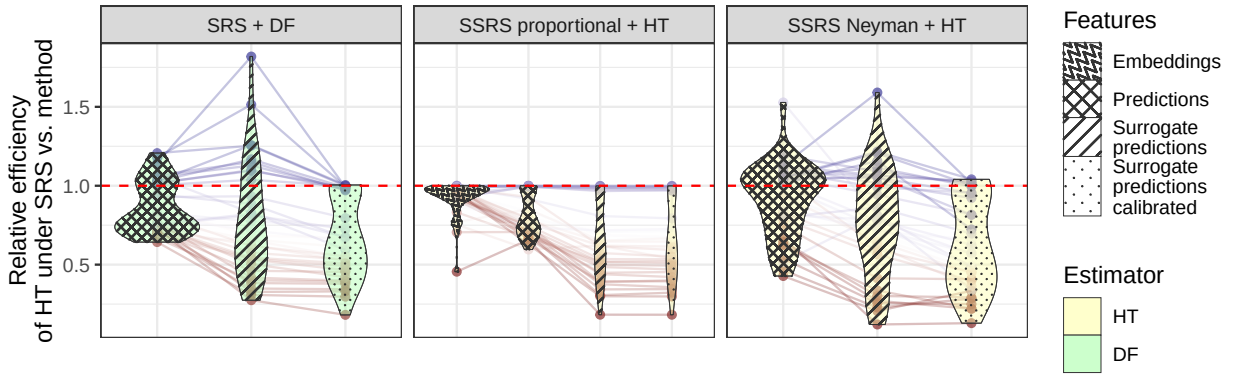

Figure 2: **Comparison of efficiency across stratification procedures, sampling designs, and estimators.** The violin plots illustrate the relative efficiency of the Horvitz-Thompson (`HT`) estimator under simple random sampling (`SRS`, red dashed line) compared to other survey sampling strategies and estimators (relative efficiency is $\mathrm{MSE}_{\pi}(\widehat{\theta}_{\mathrm{EST}})/\mathrm{MSE}_{\texttt{SRS}}(\widehat{\theta}_{\texttt{HT}})$) for estimating the accuracy of CLIP ViT-B/32 on classification tasks in the benchmark. Lower values indicate larger efficiency gains compared to the baseline. The dots and lines represent the relative efficiencies of the sampling methods and estimators on the various tasks.

Stratification. In Section 4.2, we discussed how stratifying on $\widehat{Z} = f_{\widehat{Y}}(X)$ can result in higher homogeneity within strata compared to stratifying directly on the image embeddings obtained from the same model. This is consistent with the findings in Figure 2, where the efficiency of the `HT` estimator under `SSRS` with proportional allocation is generally higher when stratification is performed on the proxy. Stratification using the proxy based on the predictions generated by a surrogate model f^* with higher performance, here CLIP ViT-L/14, additionally increases efficiency. This improvement is observed for proportional allocation across all tasks and, in most cases, for Neyman allocation as well. Calibrating these predictions does not appear to affect the formation of the strata and therefore does not affect performance under proportional allocation. However, it does change the allocation of the budget and consequently, we observed an increase in the performance of the estimates under Neyman allocation.

Estimator. The analysis in Section 4.2 suggests that, under `SRS`, the `DF` estimator has the potential to significantly improve the precision of our estimates compared to `HT`. However, as shown in Figure 2, the efficiency gains of the `DF` estimator should not be taken for granted. When $\widehat{Z}$ is based on uncalibrated model predictions, we observe that `DF` achieves higher efficiency than `HT` in many but not all of the tasks. In some cases, it performs substantially worse than `HT`. However, the `DF` estimator that leverages the calibrated proxy always achieves equal or lower MSE than HT. Consistently with our theoretical findings in Section 4.3, the values of $\mathrm{MSE}_{\texttt{SRS}}(\widehat{\theta}_{\texttt{DF}}, \widehat{\theta}_{\mathcal{D}})$ for calibrated predictions are close to those of $\mathrm{MSE}_{\texttt{SSRS},p}(\widehat{\theta}_{\texttt{HT}}, \widehat{\theta}_{\mathcal{D}})$, indicating similar gains in efficiency. Finally, as discussed in Section 4.3, `HT` and `DF` under `SSRS` yield estimates with virtually the same precision and are excluded from the figure.

Characterizing the efficiency gains. The empirical results presented so far indicate that the efficiency gains of the `DF` estimator and the stratified designs over the naive average of a completely random subset of data (i.e., `HT` under `SRS`) vary significantly across tasks. To determine when we can expect the largest gains, we turn to our theoretical analysis. In Section 4.3, we have shown that larger efficiency gains for the `DF` estimator should be expected when the MSE of $\widehat{Z}$ relative to

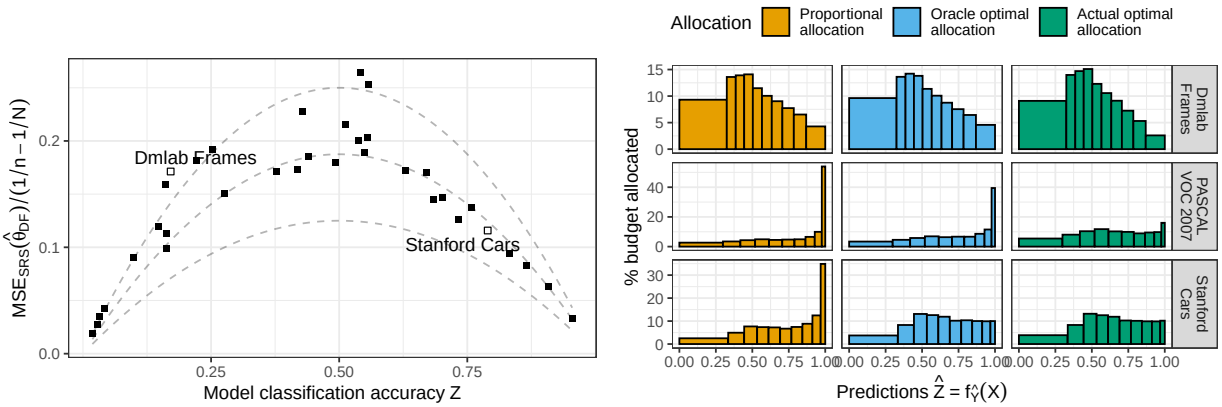

Figure 3: **Characterization of efficiency gains.** The left panel shows the mean squared error (MSE) of the difference estimator (DF) under simple random sampling (SRS, corrected by $n/(1-f)$) as a function of the zero-shot classification accuracy $N^{-1}\sum_{i\in\mathcal{D}} Z$ of CLIP ViT-B/32 evaluated on the full test sets of the LAION CLIP benchmark tasks. We construct $\widehat{Z}$ using the predictions of CLIP with ViT-B/32 as backbones. Dashed lines correspond to the relative efficiencies of 1 (highest line), 0.75, and 0.5 (lowest). In tasks where the model achieves higher classification accuracy, it also tends to have higher relative efficiency. The right panel shows the allocation of the annotation budget to each stratum through proportional and optimal (ideal based on S_{Z_h} and actual based on $\widehat{S}_{\widehat{Z}h}$) allocations across three datasets. In practice, Neyman allocation provides efficiency gains over proportional allocation only on Stanford Cars.

$\mathrm{Var}_{\mathtt{SRS}}(\widehat{\theta}_{\mathtt{HT}}) \propto \widehat{\theta}_{\mathcal{D}}(1-\widehat{\theta}_{\mathcal{D}})$ is low or similarly when $\mathbb{E}_P[\mathrm{Var}_P(Z \mid X)] \ll \mathrm{Var}_P(Z \mid X)$ in Proposition 4. Note that classifiers with the same $\mathrm{Var}_{\mathtt{SRS}}(\widehat{\theta}_{\mathtt{HT}})$ can have very different accuracy (e.g., $\widehat{\theta}_{\mathcal{D}} = 0.2$ vs. $\widehat{\theta}_{\mathcal{D}} = 0.8$) and classifiers with higher accuracy often achieve lower MSE, which will be associated with larger efficiency gains of DF over HT under SRS. This observation is confirmed by Figure 3, where we observe that $\mathrm{Var}_{\mathtt{SRS}}(\widehat{\theta}_{\mathtt{HT}})$ is similar on Dmlab Frames and Stanford Cars, but f achieves higher accuracy on the latter and also HT under SRS yields more precise estimates of $\widehat{\theta}_{\mathcal{D}}$. It is worth noting that this argument may not always hold, as a classifier's high accuracy may be explained by extreme class imbalance. Nevertheless, in the tasks we have examined, this observation generally holds. Efficiency gains of DF over HT under SRS are inherently tied to those of SSRS with proportional allocation, so similar arguments hold for that sampling design. We also mentioned in Section 4.1 that Neyman allocation may not yield sizable gains over proportional if the S_{Z_h}'s (i) are similar across strata or (ii) are poorly estimated. Figure 3 shows two examples of (i) and (ii), as well as an example where Neyman allocation leads to large gains. On Dmlab Frames, (i) occurs: The distributions of $\widehat{Z}$ conditional on $Z = 0, 1$ mostly overlap, hence proportional and Neyman allocation are similar. On Pascal VOC 2007, we observe (ii): Neyman allocates too little budget to the stratum where $\widehat{Z}$ is close to 1, which has considerable variability. Lastly, on Stanford Cars, Neyman allocation leads to a large gain, as the proportional allocation allocates too much budget to high values of $\widehat{Z}$, even though the model makes few errors in that region (i.e., mostly $Z = 1$).

6 Discussion

In this paper, we have investigated methods to evaluate the predictive performance of a machine learning model on large datasets on which only a limited amount of data can be labeled. Our findings show that, when good predictions of the model's performance are available, stratified sampling strategies and model-assisted estimators can provide more precise estimates compared to

the traditional approach of naive averaging on a data subset obtained via `SRS`.

Main takeaway. We recommend that, when selecting a data subset to annotate, CV practitioners always use stratified sampling strategies (`SSRS`) with proportional allocation, running k-means on the proxy $\widehat{Z}$ of model performance Z (Section 4.2). The more strata one can form, the higher the precision of the estimates will likely be. When the proxy $\widehat{Z}$ is well calibrated, Neyman allocation may also be used and may lead to additional efficiency gains (Section 4.1). If a data subset has already been obtained via `SRS`, then one can still leverage the `DF` estimator to increase the precision of the estimates (Section 4.3). If there is uncertainty about the quality of the proxy, the method recently proposed by [2] can be applied to adjust the extent to which the estimator relies on the proxy.

Beyond that, it is important to understand the efficiency of these estimators on out-of-distribution data. In this setting, the proxy $\widehat{Z}$ may be a poor predictor of model performance Z and consequently, the gains of `SSRS` with proportional allocation or `DF` under `SRS` relative to `HT` under `SRS` may be limited. There is also the risk that `SSRS` with Neyman allocation may yield estimates that have substantially higher variance than those obtained under proportional allocation. Therefore, caution should be exercised when using adaptive allocations and one believes that the test distribution may differ from the training distribution. These findings suggest that incorporating calibration techniques (of models and estimators) [82, 102] along with sequential sampling [105] may lead to additional improvements in the evaluation of model performance.

Acknowledgments

We thank the anonymous reviewers for their encouraging comments and valuable suggestions that have improved our manuscript. We also thank Tijana Zrnic for highlighting the connections between prediction-powered inference and our work, as well as Georgy Noarov for pointing out the link between our results and decompositions for proper scoring rules.

References

[1] Angelopoulos, A.N., Bates, S., Fannjiang, C., Jordan, M.I., Zrnic, T.: Prediction-powered inference. Science **382**(6671), 669–674 (2023)

[2] Angelopoulos, A.N., Duchi, J.C., Zrnic, T.: Ppi++: Efficient prediction-powered inference. arXiv preprint arXiv:2311.01453 (2023)

[3] Ash, J.T., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A.: Deep batch active learning by diverse, uncertain gradient lower bounds. arXiv preprint arXiv:1906.03671 (2019)

[4] Baek, C., Jiang, Y., Raghunathan, A., Kolter, J.Z.: Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. Advances in Neural Information Processing Systems **35**, 19274–19289 (2022)

[5] Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., Katz, B.: Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. Advances in neural information processing systems **32** (2019)

[6] Beery, S., Cole, E., Gjoka, A.: The iwildcam 2020 competition dataset. arXiv preprint arXiv:2004.10340 (2020)

[7] Breidt, F.J., Claeskens, G., Opsomer, J.: Model-assisted estimation for complex surveys using penalised splines. Biometrika **92**(4), 831–846 (2005)

[8] Breidt, F.J., Opsomer, J.D.: Model-Assisted Survey Estimation with Modern Prediction Techniques. Statistical Science **32**(2), 190 – 205 (2017). https://doi.org/10.1214/16-STS589

[9] Brus, D.J.: Spatial sampling with R. CRC Press (2022)

[10] Chen, M., Goel, K., Sohoni, N.S., Poms, F., Fatahalian, K., Ré, C.: Mandoline: Model evaluation under distribution shift. In: International conference on machine learning. pp. 1617–1629. PMLR (2021)

[11] Chen, T., Lumley, T.: Optimal multiwave sampling for regression modeling in two-phase designs. Statistics in medicine **39**(30), 4912–4921 (2020)

[12] Chen, T., Lumley, T.: Optimal sampling for design-based estimators of regression models. Statistics in medicine **41**(8), 1482–1497 (2022)

[13] Chen, Y., Zhang, S., Song, R.: Scoring your prediction on unseen data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 3279–3288 (June 2023)

[14] Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE **105**(10), 1865–1883 (2017)

[15] Chouldechova, A., Deng, S., Wang, Y., Xia, W., Perona, P.: Unsupervised and semi-supervised bias benchmarking in face recognition. In: European Conference on Computer Vision. pp. 289–306. Springer (2022)

[16] Chu, W., Zinkevich, M., Li, L., Thomas, A., Tseng, B.: Unbiased online active learning in data streams. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 195–203 (2011)

[17] Chuang, C.Y., Torralba, A., Jegelka, S.: Estimating generalization under distribution shifts via domain-invariant representations. arXiv preprint arXiv:2007.03511 (2020)

[18] Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., , Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2014)

[19] Clark, R.G., Steel, D.G.: Sample design for analysis using high-influence probability sampling. Journal of the Royal Statistical Society Series A: Statistics in Society **185**(4), 1733–1756 (2022)

[20] Coates, A., Ng, A., Lee, H.: An analysis of single-layer networks in unsupervised feature learning. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp. 215–223. JMLR Workshop and Conference Proceedings (2011)

[21] Cochran, W.G.: Sampling Techniques. John Wiley & Sons (1977)

[22] Cohn, D.A., Ghahramani, Z., Jordan, M.I.: Active learning with statistical models. Journal of artificial intelligence research **4**, 129–145 (1996)

[23] Deng, L.: The mnist database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine **29**(6), 141–142 (2012)

[24] Deng, W., Gould, S., Zheng, L.: What does rotation prediction tell us about classifier accuracy under varying testing environments? In: International Conference on Machine Learning. pp. 2579–2589. PMLR (2021)

[25] Deng, W., Zheng, L.: Are labels always necessary for classifier accuracy evaluation? In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15069–15078 (2021)

[26] Emma, D., Jared, J., Cukierski, W.: Diabetic retinopathy detection (2015), https://kaggle.com/competitions/diabetic-retinopathy-detection

[27] Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html

[28] Farquhar, S., Gal, Y., Rainforth, T.: On statistical bias in active learning: How and when to fix it. arXiv preprint arXiv:2101.11665 (2021)

[29] Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: 2004 conference on computer vision and pattern recognition workshop. pp. 178–178. IEEE (2004)

[30] Fuller, W.A.: Sampling Statistics. John Wiley & Sons (2011)

[31] Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. In: International conference on machine learning. pp. 1183–1192. PMLR (2017)

[32] Ganti, R., Gray, A.: Upal: Unbiased pool based active learning. In: Artificial Intelligence and Statistics. pp. 422–431. PMLR (2012)

[33] Garg, S., Balakrishnan, S., Lipton, Z.C., Neyshabur, B., Sedghi, H.: Leveraging unlabeled data to predict out-of-distribution performance. arXiv preprint arXiv:2201.04234 (2022)

[34] Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR) (2013)

[35] Graubardand, B.I., Korn, E.L.: Inference for superpopulation parameters using sample surveys. Statistical Science **17**(1), 73–96 (2002)

[36] Guillory, D., Shankar, V., Ebrahimi, S., Darrell, T., Schmidt, L.: Predicting with confidence on unseen distributions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1134–1144 (2021)

[37] Hájek, J.: Optimal strategy and other problems in probability sampling. Časopis pro pěstování matematiky **84**(4), 387–423 (1959)

[38] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

[39] Hébert-Johnson, U., Kim, M., Reingold, O., Rothblum, G.: Multicalibration: Calibration for the (computationally-identifiable) masses. In: International Conference on Machine Learning. pp. 1939–1948. PMLR (2018)

[40] Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (2019)

[41] Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., Gilmer, J.: The many faces of robustness: A critical analysis of out-of-distribution generalization. ICCV (2021)

[42] Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples. CVPR (2021)

[43] Horvitz, D.G., Thompson, D.J.: A generalization of sampling without replacement from a finite universe. Journal of the American statistical Association **47**(260), 663–685 (1952)

[44] Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: Openclip (Jul 2021). https://doi.org/10.5281/zenodo.5143773, https://doi.org/10.5281/zenodo.5143773

[45] Imberg, H., Axelson-Fisk, M., Jonasson, J.: Optimal subsampling designs. arXiv preprint arXiv:2304.03019 (2023)

[46] Imberg, H., Jonasson, J., Axelson-Fisk, M.: Optimal sampling in unbiased active learning. In: International Conference on Artificial Intelligence and Statistics. pp. 559–569. PMLR (2020)

[47] Imberg, H., Yang, X., Flannagan, C., Bärgman, J.: Active sampling: A machine-learning-assisted framework for finite population inference with optimal subsamples. arXiv preprint arXiv:2212.10024 (2022)

[48] Isaki, C.T., Fuller, W.A.: Survey design under the regression superpopulation model. Journal of the American Statistical Association **77**(377), 89–96 (1982)

[49] Jiang, Y., Nagarajan, V., Baek, C., Kolter, J.Z.: Assessing generalization of sgd via disagreement. arXiv preprint arXiv:2106.13799 (2021)

[50] Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2901–2910 (2017)

[51] Kim, M.P., Kern, C., Goldwasser, S., Kreuter, F., Reingold, O.: Universal adaptability: Target-independent inference that competes with propensity scoring. Proceedings of the National Academy of Sciences **119**(4), e2108097119 (2022)

[52] Kirsch, A., Van Amersfoort, J., Gal, Y.: Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. Advances in neural information processing systems **32** (2019)

[53] Koh, P.W., Sagawa, S., Marklund, H., Xie, S.M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R.L., Gao, I., et al.: Wilds: A benchmark of in-the-wild distribution shifts. In: International Conference on Machine Learning. pp. 5637–5664. PMLR (2021)

[54] Kossen, J., Farquhar, S., Gal, Y., Rainforth, T.: Active surrogate estimators: An active learning approach to label-efficient model evaluation. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 24557–24570. Curran Associates, Inc. (2022)

[55] Kossen, J., Farquhar, S., Gal, Y., Rainforth, T.: Active testing: Sample-efficient model evaluation. In: International Conference on Machine Learning. pp. 5753–5763. PMLR (2021)

[56] Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13). Sydney, Australia (2013)

[57] Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)

[58] Kull, M., Flach, P.: Novel decompositions of proper scoring rules for classification: Score adjustment as precursor to calibration. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part I 15. pp. 68–85. Springer (2015)

[59] LAION AI: Clip benchmark. https://github.com/LAION-AI/CLIP_benchmark

[60] LeCun, Y., Huang, F.J., Bottou, L.: Learning methods for generic object recognition with invariance to pose and lighting. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. vol. 2, pp. II–104. IEEE (2004)

[61] Lewis, D.D.: A sequential algorithm for training text classifiers: Corrigendum and additional data. In: Acm Sigir Forum. vol. 29, pp. 13–19. ACM New York, NY, USA (1995)

[62] Lewis, D.D., Catlett, J.: Heterogeneous uncertainty sampling for supervised learning. In: Machine learning proceedings 1994, pp. 148–156. Elsevier (1994)

[63] Li, Z., Ma, X., Xu, C., Cao, C., Xu, J., Lü, J.: Boosting operational dnn testing efficiency through conditioning (2019). https://doi.org/10.1145/3338906.3338930

[64] Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)

[65] Lohr, S.L.: Sampling: design and analysis. CRC press (2021)

[66] Lumley, T., Shaw, P.A., Dai, J.Y.: Connections between survey calibration estimators and semiparametric models for incomplete data. International Statistical Review **79**(2), 200–220 (2011)

[67] Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)

[68] Matthey, L., Higgins, I., Hassabis, D., Lerchner, A.: dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/ (2017)

[69] McConville, K.S., Breidt, F.J., Lee, T.C., Moisen, G.G.: Model-assisted survey regression estimation with the lasso. Journal of Survey Statistics and Methodology **5**(2), 131–158 (2017)

[70] Miller, B.A., Vila, J., Kirn, M., Zipkin, J.R.: Classifier performance estimation with unbalanced, partially labeled data. In: Torgo, L., Matwin, S., Weiss, G., Moniz, N., Branco, P. (eds.) Proceedings of The International Workshop on Cost-Sensitive Learning. Proceedings of Machine Learning Research, vol. 88, pp. 4–16. PMLR (05 May 2018)

[71] Miller, J.P., Taori, R., Raghunathan, A., Sagawa, S., Koh, P.W., Shankar, V., Liang, P., Carmon, Y., Schmidt, L.: Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In: International Conference on Machine Learning. pp. 7721–7735. PMLR (2021)

[72] Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning (2011)

[73] Neyman, J.: On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. In: Breakthroughs in Statistics: Methodology and Distribution, pp. 123–150. Springer (1992)

[74] Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian conference on computer vision, graphics & image processing. pp. 722–729. IEEE (2008)

[75] Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3498–3505. IEEE (2012)

[76] Poms, F., Sarukkai, V., Mullapudi, R.T., Sohoni, N.S., Mark, W.R., Ramanan, D., Fatahalian, K.: Low-shot validation: Active importance sampling for estimating classifier performance on rare categories. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10705–10714 (October 2021)

[77] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)

[78] Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: International conference on machine learning. pp. 5389–5400. PMLR (2019)

[79] Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Gupta, B.B., Chen, X., Wang, X.: A survey of deep active learning. ACM computing surveys (CSUR) **54**(9), 1–40 (2021)

[80] Roth, A.: Uncertain: Modern topics in uncertainty estimation (2022)

[81] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y

[82] Särndal, C.E.: The calibration approach in survey theory and practice. Survey methodology **33**(2), 99–119 (2007)

[83] Särndal, C.E., Swensson, B., Wretman, J.: Model assisted survey sampling. Springer Science & Business Media (2003)

[84] Sawade, C., Landwehr, N., Bickel, S., Scheffer, T.: Active risk estimation. In: Proceedings of the 27th International Conference on International Conference on Machine Learning. p. 951–958. ICML'10, Omnipress, Madison, WI, USA (2010)

[85] Sawade, C., Landwehr, N., Scheffer, T.: Active estimation of f-measures. In: Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., Culotta, A. (eds.) Advances in Neural Information Processing Systems. vol. 23. Curran Associates, Inc. (2010)

[86] Scheffer, T., Decomain, C., Wrobel, S.: Active hidden markov models for information extraction. In: International symposium on intelligent data analysis. pp. 309–318. Springer (2001)

[87] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5b: An open large-scale dataset for training next generation image-text models. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022), https://openreview.net/forum?id=M3Y74vmsMcY

[88] Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 (2017)

[89] Settles, B.: Active learning literature survey (2009)

[90] Siddhant, A., Lipton, Z.C.: Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. arXiv preprint arXiv:1808.05697 (2018)

[91] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 conference on empirical methods in natural language processing. pp. 1631–1642 (2013)

[92] Stallkamp, J., Schlipsing, M., Salmen, J., Igel, C.: The german traffic sign recognition benchmark: a multi-class classification competition. In: The 2011 international joint conference on neural networks. pp. 1453–1460. IEEE (2011)

[93] Taylor, J., Earnshaw, B., Mabey, B., Victors, M., Yosinski, J.: Rxrx1: An image set for cellular morphological variation across many experimental batches. In: International Conference on Learning Representations (ICLR) (2019)

[94] Tillé, Y.: Sampling and estimation from finite populations. John Wiley & Sons (2020)

[95] Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant cnns for digital pathology. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11. pp. 210–218. Springer (2018)

[96] Wald, Y., Feder, A., Greenfeld, D., Shalit, U.: On calibration and out-of-domain generalization. Advances in neural information processing systems **34**, 2215–2227 (2021)

[97] Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. In: Advances in Neural Information Processing Systems. pp. 10506–10518 (2019)

[98] Welinder, P., Welling, M., Perona, P.: A lazy man's approach to benchmarking: Semisupervised classifier evaluation and recalibration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2013)

[99] Wenzel, F., Dittadi, A., Gehler, P., Simon-Gabriel, C.J., Horn, M., Zietlow, D., Kernert, D., Russell, C., Brox, T., Schiele, B., et al.: Assaying out-of-distribution generalization in transfer learning. Advances in Neural Information Processing Systems **35**, 7181–7198 (2022)

[100] Wu, C., Sitter, R.R.: A model-calibration approach to using complete auxiliary information from survey data. Journal of the American Statistical Association **96**(453), 185–193 (2001)

[101] Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 3485–3492 (June 2010). https://doi.org/10.1109/CVPR.2010.5539970

[102] Yu, Y., Bates, S., Ma, Y., Jordan, M.: Robust calibration with multi-domain temperature scaling. Advances in Neural Information Processing Systems **35**, 27510–27523 (2022)

[103] Yu, Y., Yang, Z., Wei, A., Ma, Y., Steinhardt, J.: Predicting out-of-distribution error with the projection norm. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 25721–25746. PMLR (17–23 Jul 2022)

[104] Zhai, X., Puigcerver, J., Kolesnikov, A., Ruyssen, P., Riquelme, C., Lucic, M., Djolonga, J., Pinto, A.S., Neumann, M., Dosovitskiy, A., Beyer, L., Bachem, O., Tschannen, M., Michalski, M., Bousquet, O., Gelly, S., Houlsby, N.: The visual task adaptation benchmark (2020), https://openreview.net/forum?id=BJena3VtwS

[105] Zrnic, T., Candès, E.J.: Active statistical inference. arXiv preprint arXiv:2403.03208 (2024)

[106] Zrnic, T., Candès, E.J.: Cross-prediction-powered inference. Proceedings of the National Academy of Sciences **121**(15), e2322083121 (2024)

Appendix

This appendix complements our main paper "A Framework for Efficient Model Evaluation through Stratification, Sampling, and Estimation."

Organization

The appendix is organized as follows.

- In Appendix A, we provide proofs of the theoretical results presented in the paper. Specifically, this section includes the following proofs:
  - Proof of Proposition 1 (Appendix A.1).
  - Proof of Proposition 2 (Appendix A.2).
  - Proof of Proposition 4 (Appendix A.3).
- In Appendix B, we present additional results that complement the findings in the paper. Specifically, this section includes the following results:
  - Breakdown of the results shown in Figure 2 (Appendix B.1).
  - Comparison of different methods for estimating classifiers mean squared error (MSE) and cross-entropy (Appendix B.2).
  - Assessment of classification accuracy in CLIP models using linear probing (Appendix B.3).
  - Tests on CLIP models with ResNet and ConvNeXT visual encoders (Appendix B.4).
  - Further information on the out-of-distribution results presented in Figure 7 (Appendix B.5).

A Proofs

A.1 Proof of Proposition 1

This proof is standard and can be found in survey sampling textbooks [21, 94]. For the reader's convenience, we provide the proof below in our notation.

Part 1. To prove that $\mathrm{MSE}_{\mathtt{SSRS},p}(\widehat{\theta}_{\mathtt{HT}}, \widehat{\theta}_{\mathcal{D}}) \leq \mathrm{MSE}_{\mathtt{SRS}}(\widehat{\theta}_{\mathtt{HT}}, \widehat{\theta}_{\mathcal{D}})$, recall that $\mathrm{MSE}_{\mathtt{SSRS}}(\widehat{\theta}_{\mathtt{HT}}, \widehat{\theta}_{\mathcal{D}}) = \mathrm{Var}_{\mathtt{SSRS},p}(\widehat{\theta}_{\mathtt{HT}})$ and $\mathrm{MSE}_{\mathtt{SRS}}(\widehat{\theta}_{\mathtt{HT}}, \widehat{\theta}_{\mathcal{D}}) = \mathrm{Var}_{\mathtt{SRS}}(\widehat{\theta}_{\mathtt{HT}})$ as the estimators are unbiased with respect to the sampling design. Thus, we need to show that $\mathrm{Var}_{\mathtt{SRS}}(\widehat{\theta}_{\mathtt{HT}}) - \mathrm{Var}_{\mathtt{SSRS},p}(\widehat{\theta}_{\mathtt{HT}}) \geq 0$. We can rewrite

$$(N-1)S_Z^2 = \sum_{h=1}^{H}(N_h - 1)S_{Z_h}^2 + \sum_{h=1}^{H} N_h(\widehat{\theta}_{\mathcal{D}_h} - \widehat{\theta}_{\mathcal{D}})^2,$$

where $\mathcal{D}_h = N_h^{-1}\sum_{i\in\mathcal{D}_h} Z_i$. When $(N_h - 1)/N \approx N_h/N$,

$$S_{Z_h}^2 \approx \sum_{h=1}^{H}\frac{N_h}{N}S_{Z_h}^2 + \sum_{h=1}^{H}\frac{N_h}{N}(\widehat{\theta}_{\mathcal{D}_h} - \widehat{\theta}_{\mathcal{D}})^2.$$

Consequently, we have

$$\mathrm{Var}_{\mathtt{SRS}}(\widehat{\theta}_{\mathtt{HT}}) - \mathrm{Var}_{\mathtt{SSRS},p}(\widehat{\theta}_{\mathtt{HT}}) \approx \frac{1-f}{n}\sum_{h=1}^{H}\frac{N_h}{N}(\widehat{\theta}_{\mathcal{D}_h} - \widehat{\theta}_{\mathcal{D}})^2,$$

completing the first part of the proof.

Part 2. To show that $\mathrm{MSE}_{\mathtt{SSRS},o}(\widehat{\theta}_{\mathtt{HT}}, \widehat{\theta}_{\mathcal{D}}) \leq \mathrm{MSE}_{\mathtt{SSRS},p}(\widehat{\theta}_{\mathtt{HT}}, \widehat{\theta}_{\mathcal{D}})$, and equivalently $\mathrm{Var}_{\mathtt{SSRS},o}(\widehat{\theta}_{\mathtt{HT}}) \leq \mathrm{Var}_{\mathtt{SSRS},p}(\widehat{\theta}_{\mathtt{HT}})$, we observe that

$$\begin{aligned}\mathrm{Var}_{\mathtt{SSRS},p}(\widehat{\theta}_{\mathtt{HT}}) - \mathrm{Var}_{\mathtt{SSRS},o}(\widehat{\theta}_{\mathtt{HT}}) &= \frac{1}{n}\sum_{h=1}^{H}\frac{N_h}{N}S_{Z_h}^2 - \frac{1}{n}\Big(\sum_{h=1}^{H}\frac{N_h}{N}S_{Z_h}\Big)^2 \\ &= \frac{1}{n}\sum_{h=1}^{H}\frac{N_h}{N}(S_{Z_h} - \overline{S_Z})^2,\end{aligned}$$

where $\overline{S_Z} = \sum_{h=1}^{H}(N_h/N)S_{Z_h}$. This completes the second part of the proof.

A.2 Proof of Proposition 2

Note: Analogous result can be also be derived (in more generality) by using the three-term decomposition of proper scoring rules in [58, Section 5]. Below we provide a proof using our notation for the setting of this paper.

Recall that the expected variance conditional on $X = (X_1, \ldots, X_N)$ of the HT estimator under SSRS with proportional allocation is

$$\mathbb{E}_P[\mathrm{Var}_{\mathtt{SSRS},p}(\widehat{\theta}_{\mathtt{HT}}) \,|\, X] = \frac{1-f}{n}\sum_{h=1}^{H}\frac{N_h}{N}\frac{1}{N_h-1}\sum_{i\in\mathcal{D}_h}\mathbb{E}[(Z_i - \widehat{\theta}_{\mathcal{D}_h})^2 \,|\, X].$$

Now, let $\widehat{Z} = \mathbb{E}_P[Z \,|\, X]$ and $\widehat{\overline{Z}}_{\mathcal{D}_h} = N_h^{-1}\sum_{i\in\mathcal{D}_h}\widehat{Z}_i$. We can decompose $(Z_i - \widehat{\theta}_{\mathcal{D}_h})^2$ as follows:

$$\begin{aligned}(Z_i - \widehat{\theta}_{\mathcal{D}_h})^2 &= (Z_i - \widehat{Z}_i + \widehat{Z}_i - \widehat{\theta}_{\mathcal{D}_h})^2 \\ &= (Z_i - \widehat{Z}_i)^2 + (\widehat{Z}_i - \widehat{\theta}_{\mathcal{D}_h})^2 + 2(Z_i - \widehat{Z}_i)(\widehat{Z}_i - \widehat{\theta}_{\mathcal{D}_h}) \\ &= (Z_i - \widehat{Z}_i)^2 + (\widehat{Z}_i - \widehat{\overline{Z}}_{\mathcal{D}_h} + \widehat{\overline{Z}}_{\mathcal{D}_h} - \widehat{\theta}_{\mathcal{D}_h})^2 + 2(Z_i - \widehat{Z}_i)(\widehat{Z}_i - \widehat{\theta}_{\mathcal{D}_h}) \\ &= (Z_i - \widehat{Z}_i)^2 + (\widehat{Z}_i - \widehat{\overline{Z}}_{\mathcal{D}_h})^2 + (\widehat{\overline{Z}}_{\mathcal{D}_h} - \widehat{\theta}_{\mathcal{D}_h})^2 \\ &\quad + 2(\widehat{Z}_i - \widehat{\overline{Z}}_{\mathcal{D}_h})(\widehat{\overline{Z}}_{\mathcal{D}_h} - \widehat{\theta}_{\mathcal{D}_h}) + 2(Z_i - \widehat{Z}_i)(\widehat{Z}_i - \widehat{\theta}_{\mathcal{D}_h}).\end{aligned}$$

We can show that

$$\sum_{i\in\mathcal{D}_h}(\widehat{Z}_i - \widehat{\overline{Z}}_{\mathcal{D}_h})(\widehat{\overline{Z}}_{\mathcal{D}_h} - \widehat{\theta}_{\mathcal{D}_h}) = (\widehat{\overline{Z}}_{\mathcal{D}_h} - \widehat{\theta}_{\mathcal{D}_h})\sum_{i\in\mathcal{D}_h}(\widehat{Z}_i - \widehat{\overline{Z}}_{\mathcal{D}_h}) = 0.$$

Then, since

$$\frac{1}{N_h-1}\sum_{i\in\mathcal{D}_h}[(\widehat{\theta}_{\mathcal{D}_h} - \widehat{\overline{Z}}_{\mathcal{D}_h})^2 + 2(Z_i - \widehat{Z}_i)(\widehat{Z}_i - \widehat{\theta}_{\mathcal{D}_h})] \approx (\widehat{\theta}_{\mathcal{D}_h}^2 - \widehat{\overline{Z}}_{\mathcal{D}_h}^2) + 2\widehat{Z}_i(Z_i - \widehat{Z}_i),$$

when $1/N_h \approx 0$, we obtain (7). However, we will continue the proof without this assumption.

By the assumption of independence, we also have

$$\begin{aligned}
&\mathbb{E}_P[(Z_i - \widehat{Z}_i)(\widehat{Z}_i - \widehat{\theta}_{\mathcal{D}_h}) \,|\, X] \\
&= -\mathbb{E}_P[(Z_i - \mathbb{E}_P[Z_i \,|\, X_i])(\widehat{\theta}_{\mathcal{D}_h} - \mathbb{E}_P[Z_i \,|\, X_i]) \,|\, X] \\
&= -\frac{1}{N_h}\mathbb{E}_P[(Z_i - \mathbb{E}_P[Z_i \,|\, X_i])(Z_i - \mathbb{E}_P[Z_i \,|\, X_i]) \,|\, X] \\
&= -\frac{1}{N_h}\mathrm{Var}_P(Z_i \,|\, X_i).
\end{aligned}$$

Using similar arguments, we obtain

$$\mathbb{E}_P[(\widehat{\overline{Z}}_{\mathcal{D}_h} - \widehat{\theta}_{\mathcal{D}_h})^2 \,|\, X] = \frac{1}{N_h^2}\sum_{i\in\mathcal{D}_h}\mathrm{Var}_P(Z_i \,|\, X_i).$$

Thus, we have

$$\begin{aligned}
&\sum_{i\in\mathcal{D}_h}\mathbb{E}_P[(Z_i - \widehat{\theta}_{\mathcal{D}_h})^2 \,|\, X] \\
&= \sum_{i\in\mathcal{D}_h}\mathbb{E}_P[(Z_i - \widehat{Z}_i)^2 \,|\, X] + \sum_{i\in\mathcal{D}_h}(\widehat{Z}_i - \widehat{\overline{Z}}_{\mathcal{D}_h})^2 - \frac{1}{N_h}\sum_{i\in\mathcal{D}_h}\mathrm{Var}_P(Z_i \,|\, X_i) \\
&= \sum_{i\in\mathcal{D}_h}\mathrm{Var}_P(Z_i \,|\, X_i) + \sum_{i\in\mathcal{D}_h}(\widehat{Z}_i - \widehat{\overline{Z}}_{\mathcal{D}_h})^2 - \frac{1}{N_h}\sum_{i\in\mathcal{D}_h}\mathrm{Var}_P(Z_i \,|\, X_i).
\end{aligned}$$

Finally, the above implies that

$$\begin{aligned}
&\mathbb{E}_P[\mathrm{Var}_{\mathtt{SSRS},p}(\widehat{\theta}_{\mathtt{HT}}) \,|\, X] \\
&= \frac{1-f}{n}\sum_{h=1}^{H}\frac{N_h}{N(N_h-1)}\sum_{i\in\mathcal{D}_h}\left\{\mathrm{Var}_P(Z_i \,|\, X_i) + (\widehat{Z}_i - \widehat{\overline{Z}}_{\mathcal{D}_h})^2 - \frac{1}{N_h}\mathrm{Var}_P(Z_i \,|\, X_i)\right\} \\
&= \frac{1-f}{n}\sum_{h=1}^{H}\frac{N_h}{N(N_h-1)}\sum_{i\in\mathcal{D}_h}\left\{\frac{N_h-1}{N_h}\mathrm{Var}_P(Z_i \,|\, X_i) + (\widehat{Z}_i - \widehat{\overline{Z}}_{\mathcal{D}_h})^2\right\} \\
&= \frac{1-f}{n}\frac{1}{N}\left\{\sum_{i\in\mathcal{D}}\mathrm{Var}_P(Z_i \,|\, X_i) + \sum_{h=1}^{H}\frac{N_h}{N_h-1}\sum_{i\in\mathcal{D}_h}(\widehat{Z}_i - \widehat{\overline{Z}}_{\mathcal{D}_h})^2\right\} \\
&= \frac{1-f}{n}\frac{1}{N}\sum_{i\in\mathcal{D}}\mathrm{Var}_P(Z_i \,|\, X_i) + \frac{1-f}{n}\sum_{h=1}^{H}\frac{N_h}{N}S^2_{\widehat{Z}h}.
\end{aligned}$$

Note that the first term does not depend on the specific strata, hence the stratification procedure only affects the second term. This completes the proof.

A.3 Proof of Proposition 4

We start with the decomposition in (8):

$$\mathrm{Var}_{\mathtt{SRS}}(\widehat{\theta}_{\mathtt{DF}}) = \frac{1-f}{n}\left\{\frac{1}{N-1}\sum_{i=1}^{N}(Z_i - \widehat{Z}_i)^2 - \frac{N}{N-1}(\widehat{\theta}_{\mathcal{D}} - \widehat{\overline{Z}})^2\right\}. \tag{10}$$

By the independence of $\widehat{Z}_i$ and Z_i, we have

$$\begin{aligned} N^2\mathbb{E}_P[(\widehat{\theta}_{\mathcal{D}} - \widehat{\overline{Z}})^2 \,|\, X] &= \sum_{i=1}^{N} \mathbb{E}_P[(Z_i - \widehat{Z}_i)^2 \,|\, X] + \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} \mathbb{E}_P[Z_i - \widehat{Z}_i \,|\, X] \cdot \mathbb{E}_P[Z_j - \widehat{Z}_j \,|\, X] \\ &= \sum_{i=1}^{N} \mathbb{E}_P[(Z_i - \widehat{Z}_i)^2 \,|\, X]. \end{aligned}$$

Therefore, we obtain

$$\mathbb{E}_P[\mathrm{Var}_{\mathtt{SRS}}(\widehat{\theta}_{\mathtt{DF}})] = \frac{1-f}{n} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_P[\mathbb{E}_P[(Z_i - \widehat{Z}_i)^2 \,|\, X]] = \frac{1-f}{n} \mathbb{E}[\mathrm{Var}_P(Z \,|\, X)].$$

The remaining part of the proof is straightforward and is thus omitted.

B Extended Results and Analyses

B.1 Detailed Analysis of Main Results in Section 5

The datasets and tasks included in our experiments of Section 5, together with the efficiency of HT under simple random sampling relative to other methods, are listed in Table 1.

Table 1: Breakdown of results in Figure 2. Each number corresponds to the relative efficiency of model accuracy estimates obtained through different sampling designs and estimators compared to HT under SRS, the Horvitz-Thompson estimator under simple random sampling. The proxy $\widehat{Z}$ is constructed using model predictions f, surrogate model predictions f^*, and calibrated predictions of a surrogate model on in-distribution data f^{*c}. Note that "emb" refers to the embeddings in the table.

Datasets	**Methods**									
	SRS + DF			SSRS, p + HT				SSRS, o + HT		
Dataset Name and Reference	f	f^*	f^{*c}	emb	f	f^*	f^{*c}	f	f^*	f^{*c}
Caltech 101 [29]	0.66	0.66	0.54	0.75	0.57	0.60	0.54	0.47	0.81	0.33
Stanford Cars [56]	0.69	0.35	0.34	0.98	0.69	0.34	0.33	0.51	0.24	0.23
CIFAR-10 [57]	0.77	0.45	0.38	0.95	0.74	0.41	0.38	0.60	0.35	0.20
CIFAR-100 [57]	0.69	0.50	0.46	0.92	0.67	0.47	0.46	0.75	0.52	0.37
CLEVR (distance) [50]	1.16	1.47	1.01	0.99	0.97	0.99	1.01	1.11	1.08	1.03
CLEVR (count) [50]	0.73	0.63	0.50	0.77	0.67	0.52	0.50	0.78	0.63	0.44
Describable Text Features [18]	0.82	0.71	0.65	0.95	0.81	0.64	0.65	1.15	0.87	0.69
DR Detection [26]	1.02	1.11	0.97	0.96	1.00	0.97	0.97	1.04	1.03	0.95
DMLab Frames [104]	1.22	1.26	0.99	0.99	1.00	1.01	0.98	1.09	1.55	1.01
dSprites (orientation) [68]	1.06	1.21	0.95	0.93	0.98	1.02	0.95	1.13	1.05	0.89
dSprites (x position) [68]	1.03	1.04	0.97	1.00	1.02	1.01	0.97	1.08	1.24	1.01
dSprites (y position) [68]	1.12	1.82	0.99	1.01	0.97	1.00	0.99	1.03	1.01	0.96
EuroSAT [40]	0.85	0.69	0.65	0.75	0.84	0.66	0.66	0.95	0.74	0.63
FGVC aircraft [67]	0.83	0.76	0.71	0.91	0.80	0.68	0.70	0.80	0.73	0.63
Oxford 102 Flower [74]	0.68	0.41	0.38	0.94	0.67	0.40	0.38	0.64	0.32	0.28
GTSRB [92]	0.72	0.46	0.47	0.71	0.72	0.45	0.47	0.78	0.41	0.43
ImageNet-A [42]	1.06	0.81	0.60	0.99	0.95	0.59	0.60	1.38	0.82	0.50
ImageNet-R [41]	0.63	0.31	0.30	0.96	0.61	0.30	0.30	0.58	0.28	0.21
ImageNet-1K [81]	0.74	0.54	0.51	0.97	0.73	0.51	0.51	0.92	0.67	0.46
ImageNet Sketch [97]	0.72	0.54	0.49	0.97	0.70	0.50	0.50	0.88	0.63	0.45
ImageNetV2 [78]	0.75	0.58	0.52	0.97	0.74	0.54	0.52	0.97	0.70	0.50
KITTI Distance [34]	1.06	1.12	0.92	0.69	0.97	0.93	0.88	1.09	1.04	0.95
MNIST [23]	0.73	0.27	0.19	0.47	0.66	0.18	0.19	0.67	0.11	0.12
ObjectNet [5]	0.75	0.51	0.41	0.96	0.72	0.45	0.41	1.06	0.60	0.35
Oxford-IIIT Pet [75]	0.73	0.36	0.41	0.98	0.73	0.35	0.42	0.46	0.19	0.20
PASCAL VOC 2007 [27]	0.76	0.85	0.72	0.88	0.75	0.74	0.72	1.07	1.14	0.71
PCam [95]	1.03	1.17	0.99	0.91	0.99	0.98	0.99	1.05	1.05	1.03
Rendered SST-2 [91]	1.06	1.14	0.98	0.98	1.00	0.98	0.98	1.11	1.15	1.02
NWPU-RESISC45 [14]	0.80	0.63	0.55	0.96	0.78	0.58	0.55	0.97	0.77	0.49
SmallNorb (Azimuth) [60]	0.97	1.24	1.06	1.03	0.93	0.98	1.05	0.99	1.14	1.11
smallNORB (Elevation) [60]	0.99	1.09	1.04	0.99	0.97	0.97	1.03	1.01	1.09	1.07
STL-10 [20]	0.75	0.31	0.34	0.95	0.71	0.29	0.33	0.43	0.19	0.23
SUN397 [101]	0.77	0.66	0.61	0.99	0.77	0.62	0.61	0.85	0.72	0.58
Street View House Numbers [72]	0.76	0.81	0.68	0.77	0.75	0.66	0.68	0.83	0.87	0.71

B.2 Experiments on Other Classification Metrics

In the following suite of experiments, we consider the following evaluation metrics for the predictions $f(X) = (f_1(X), \ldots, f_K(X))$ made by the classifier f:

- Mean squared error (MSE), where $Z = (1 - f_Y(X))^2$. The expected value of Z given X is $\mathbb{E}_P[Z \mid X] = \sum_{k=1}^{K} \mathbb{P}_P(Y = k \mid X)(1 - f_k(X))^2$.
- Cross-entropy loss, where $Z = -\log f_Y(X)$. The expected value of Z given X is $\mathbb{E}[Z \mid X] = -\sum_{k=1}^{K} \mathbb{P}_P(Y = k \mid X) \log f_k(X)$.

To estimate $S^2_{Z_h}$, which is needed for allocating the budget to strata under Neyman allocation (i.e., set n_h), we use the plug-in estimator $\widehat{S}^2_{Z_h} = \frac{1}{N_h} \sum_{i \in \mathcal{D}_h} \widehat{Z}_i^{(2)} - \widehat{\overline{Z}}^{(2)}_{\mathcal{D}_h}$ where $\widehat{Z}_i^{(2)}$ is an estimator of $\mathbb{E}[Z_i^2 \mid X_i]$ and $\widehat{\overline{Z}}^{(2)}_{\mathcal{D}_h}$ is its empirical average taken over $\mathcal{D}_h$.

Figure 4 shows the results obtained using the same setup and models as in Section 5. Similar to the accuracy analysis, we observe that the proportional allocation estimates made by stratifying over $\widehat{Z}$ using ViT-L/14's predictions generally outperform those using CLIP ViT-B/32's predictions, which in turn are more precise than those made by stratifying on CLIP ViT-B/32 embeddings. Estimates from proportional allocation are more accurate than those from Neyman allocation on some tasks where Neyman sometimes underperforms compared to the baseline. However, proportional allocation does not achieve the substantial improvements seen with Neyman's on other tasks. The `DF` estimator performs better than the baseline on some tasks but worse on others. In additional experiments we found that using a $\widehat{Z}$ that is trained on in-distribution validation data boosts the performance of both `DF` and Neyman, allowing them to always improve upon the baseline. This is consistent with the findings in Figure 2, where the lack of calibration in predictions can lead to larger variances compared to the baseline. Overall, each method significantly reduces the error in estimating model MSE and cross-entropy loss.

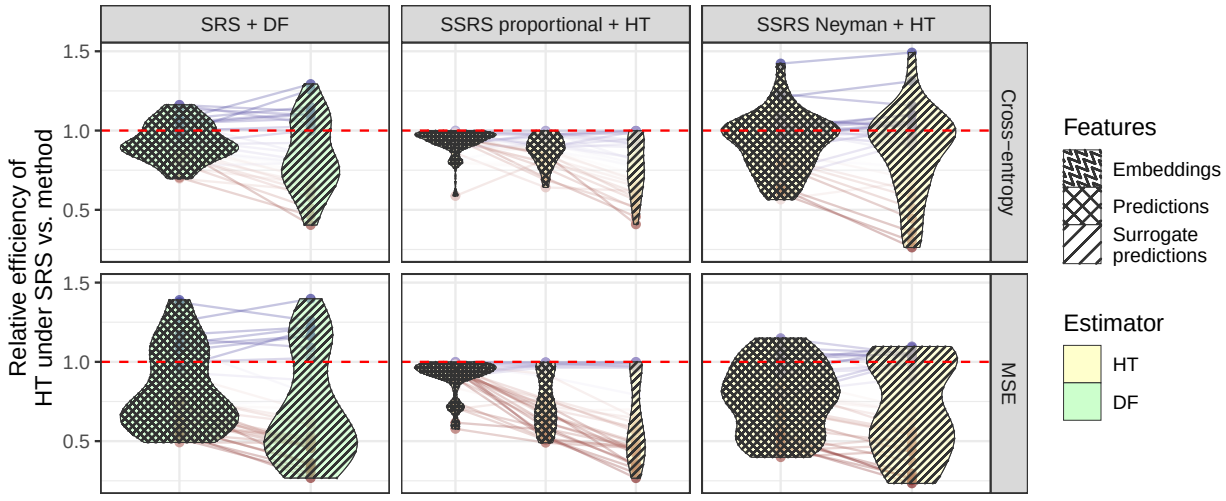

Figure 4: **Comparison of efficiency across stratification procedures, sampling designs, and estimators for estimating** MSE **and cross-entropy.** We evaluate the zero-shot accuracy of CLIP ViT-B/32 and generate surrogate predictions using CLIP ViT-L/14, also in the zero-shot setting. For more details, refer to Figure 2.

B.3 Experiments on CLIP Models with Linear Probing

We compare the efficiency of the methods in estimating the binary classification accuracy of predictions made by CLIP ViT-B/32 and CLIP ViT-L/14 using linear probing. In this setup, the model embeddings are frozen and a single linear layer is trained on top of them. We train it using the LAION CLIP repository code with the default data splits [59]. Across the tasks evaluated in the zero-shot setting and with linear probing, the latter consistently achieves higher accuracy compared to the zero-shot setting.

The main results from this set of experiments are shown in Figure 5. For easy comparison, we also report the efficiency of the methods for CLIP ViT-B/32 in the zero-shot setting. In contrast to Figure 2, we observe that with linear probing, most of the methods outperform the baseline of `HT` under `SRS`. However, in the zero-shot setting, the methods tend to perform worse. This could be attributed to the lower MSE achieved by training the linear layer, as discussed in Section 4. Consistent with our findings in Section 5, calibration of the proxy $\widehat{Z}$ (based only on f) improves efficiency for `HT` under Neyman allocation and for `DF` under `SSRS`, but not for `HT` under proportional allocation. In addition, the differences in efficiency between ViT-B/32 and ViT-L/14 with linear probing become less pronounced compared to the zero-shot setting.

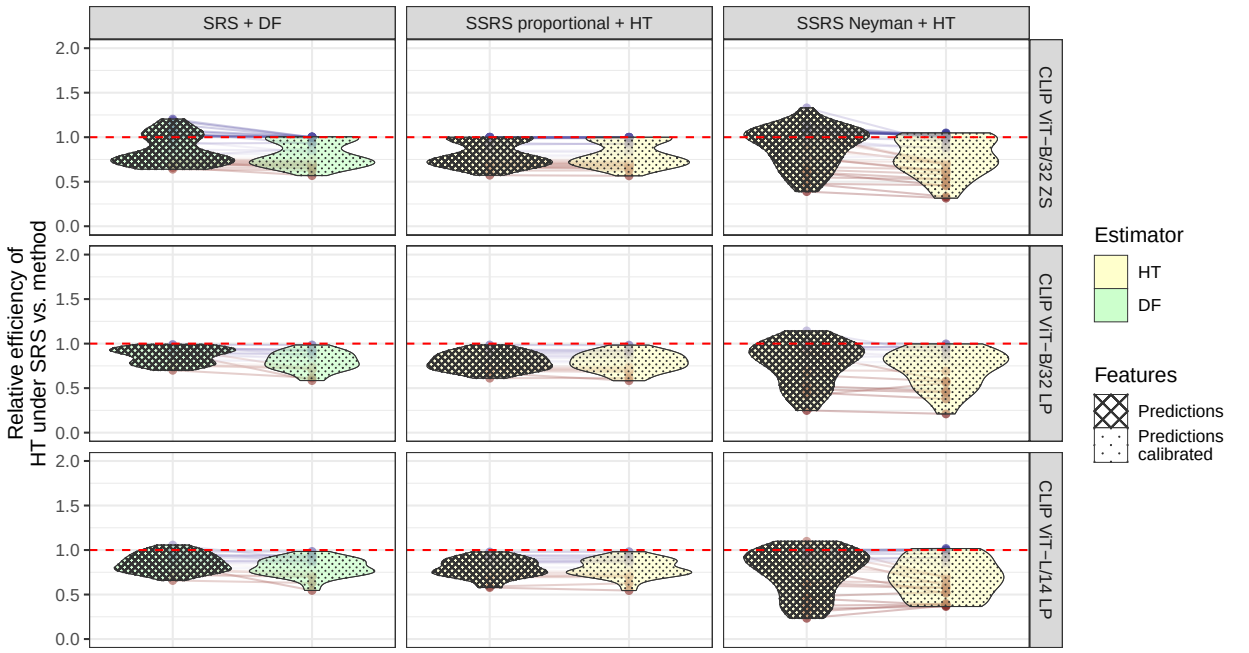

Figure 5: **Comparison of efficiency across sampling designs, estimators, and CLIP models in the zero-shot setting (ZS) and with linear probing (LP).** In this figure, we present the results specifically for the proxy $\widehat{Z}$ of Z built on the model being evaluated. For a more detailed explanation of the figure, please see Figure 2.

B.4 Experiments on Other Visual Encoders of CLIP Models

We compare the methods in estimating the zero-shot classification accuracy of CLIP models with ResNet 50 [38] and ConvNeXT base [64] as visual encoders. We obtain the surrogate predictions using ResNet 101 and ConvNeXT XXLarge respectively.

The results are shown in Figure 6. At a high level, the takeaways in Section 5 hold in this context as well. More specifically, using `SSRS` with proportional allocation always lowers the variance of the estimates of model accuracy compared to using `HT` under `SRS`. The stratification based on the predictions is more effective than the one on the embeddings. Similarly to Figure 2, the efficiency of `DF` under `SRS` and of `HT` under Neyman allocation varies across datasets and is not always superior to the baseline. Calibration, however, improves efficiency across most datasets. As noted previously, we also find that leveraging surrogate predictions from models with higher accuracy typically enhances the precision of our estimates for these architectures as well. Lastly, we find that ConvNeXT achieves far higher performance in the classification tasks compared to ResNet and the efficiency gains over `HT` under `SRS` for the former are consistently larger across all methods.

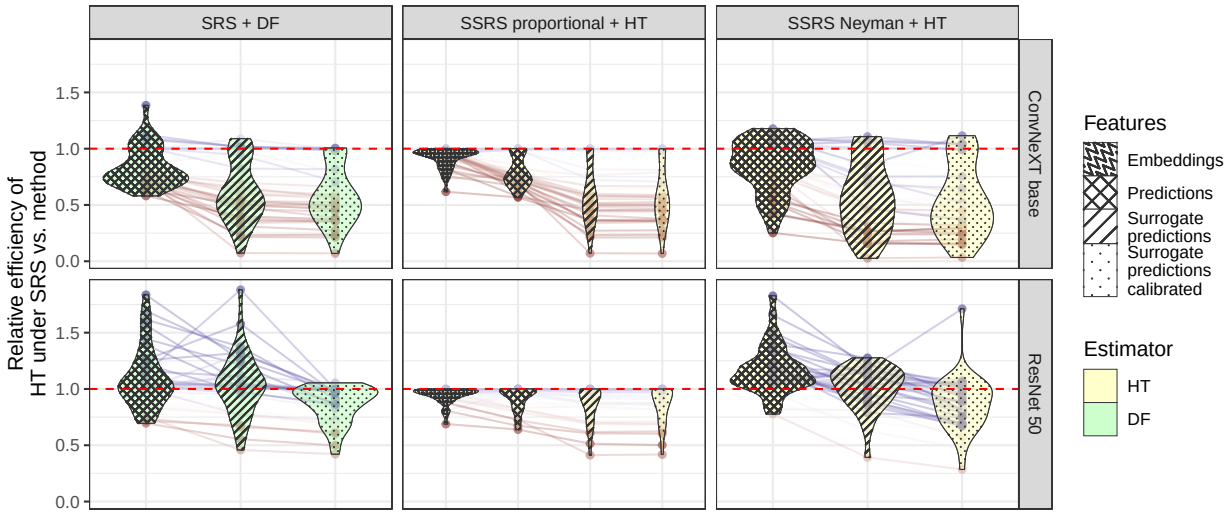

Figure 6: **Comparison of efficiency across sampling designs, estimators, and models.** We evaluate the performance of ResNet 50 on the LAION CLIP benchmark tasks using surrogate predictions from ResNet 101. Both models are pretrained on the same data. Similarly, for ConvNeXT, we assess the accuracy of the base model using surrogate predictions from ConvNeXT XXLarge. Please see Figure 2 for a detailed explanation of the elements in the figure.

B.5 Comparison of In- versus Out-of-Distribution Data

To evaluate the performance of our methods on in- vs. out-of-distribution data, we finetune a ResNet 18 model on the RxRx1 [93] and iWildCam [6] datasets from the WILDS out-of-distribution benchmark [53]. This is done using SGD on the official train splits of the datasets. We then calibrate the models using the in-distribution validation split, and evaluate their performance on in- and out-of-distribution test domains. In Figure 7, we compare the performance of the in-distribution and out-of-distribution settings. The figure highlights that when estimating model performance, efficiency gains from stratified sampling procedures are likely to be higher on the in-distribution data.

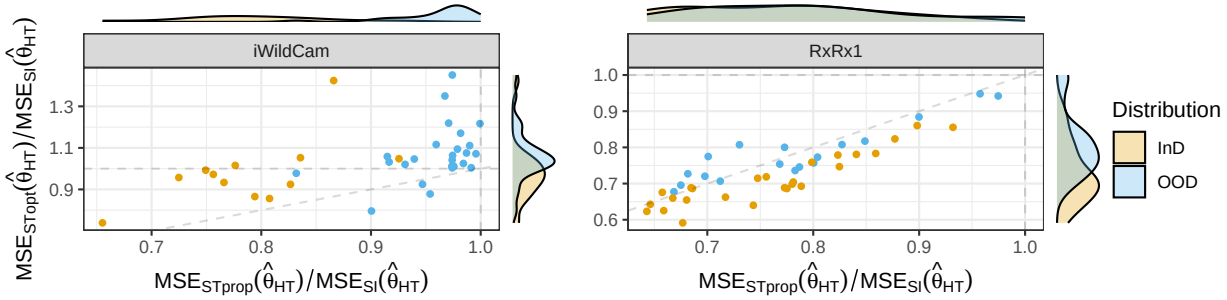

Figure 7: **Comparison of the efficiency of sampling designs and estimators on in-distribution versus out-of-distribution data.** The relative efficiencies of `HT` under `SRS` vs. the `HT` estimator under `SSRS` with proportional allocation (horizontal axis) and Neyman allocation (vertical axis) are shown in the plot. The methods estimate the classification accuracy of a Resnet 18 model trained and evaluated on the WILDS-iWildCam and WILDS-RxRx1 datasets. Stratification is done on $\widehat{Z}$ using the predictions of Z made by the models. Each point in the plot represents one domain in the datasets, with kernel density estimates of these points shown on the margins. We observe that the methods perform better compared to the baseline when the model is evaluated on in-distribution data. On the out-of-distribution data of the iWildCam dataset, Neyman allocation generally performs worse than proportional allocation and often worse than `SRS`.