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 Section 6, we provide a review of related work.
- In Section 7, we provide proofs of the theoretical results presented in the paper. Specifically, this section includes the following proofs:
  - Proof of Proposition 1 (Section 7.1).
  - Proof of Proposition 2 (Section 7.2).
  - Proof of Proposition 3 (Section 7.3).
- In Section 8, 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 (Section 8.1).
  - Comparison of different methods for estimating classifiers mean squared error (MSE) and cross-entropy (Section 8.2).
  - Assessment of classification accuracy in CLIP models using linear probing (Section 8.3).
  - Tests on CLIP models with ResNet and ConvNeXT visual encoders (Section 8.4).
  - Further information on the out-of-distribution results presented in Figure 7 (Section 8.5).

6 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.

6.1 Related Work in Survey Sampling

The question of efficient or precise evaluation is essentially analogous to problems encountered in survey sampling [21, 30, 35, 48, 65, 83] 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, 82, 83, 100]. Our findings align with existing survey sampling literature [7, 8, 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 [11,12,19,37,45]. 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 [9, 94]. 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].

6.2 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 [47, 63]. The devised sampling designs generally rely on stratification or unequal probability sampling, using predictions of model accuracy generated by the model itself [76, 84, 85] or by a surrogate model [54, 55].

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 4).

Model performance estimation on unlabeled or partially labeled data. Our paper is related to efforts on the estimation of model performance on unlabeled data [10,13,15,24,71,99,103]. 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 [33,36], and disagreement between the predictions made by multiple similar models [4,17,49]. 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 [39,51,80]. Some studies also address this challenge using a mix of unlabeled and labeled data, applying parametric models to predictions and existing labels [70,98]. 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 [16,22,28,32,61, 89,90]. 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 [31,79,88] and batch sampling strategies [3,52]. 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.

7 Proofs

7.1 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

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

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

7.2 Proof of Proposition 2

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

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

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

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

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

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

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.

7.3 Proof of Proposition 3

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

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

Therefore, we obtain

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

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

8 Extended Results and Analyses

8.1 Detailed Analysis of Main Results in Section 4

The datasets and tasks included in our experiments of Section 4, 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

8.2 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 4. 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.

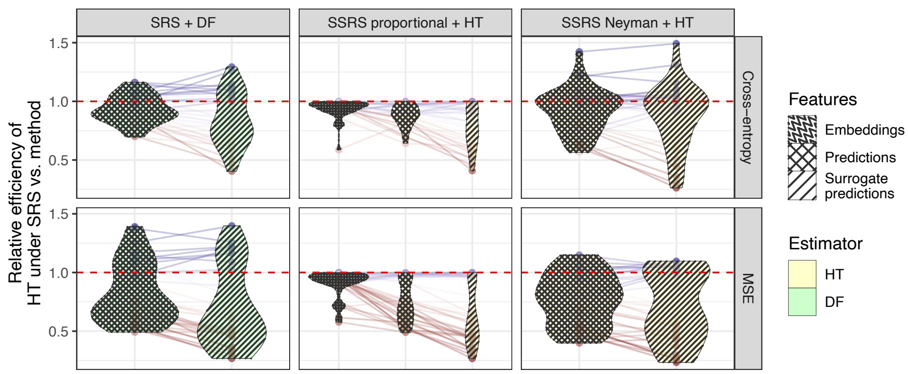

Fig. 4: Comparison of efficiency across stratification procedures, sampling designs, and estimators for estimating MSE **and cross-entropy.** We evaluate the zero-shot accuracy of CLIP ViT-B/32 and generate surrogate predictions using CLIP ViT-L/14, also in the zero-shot setting. For more details, refer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

8.3 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 3. Consistent with our findings in Section 4, 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.

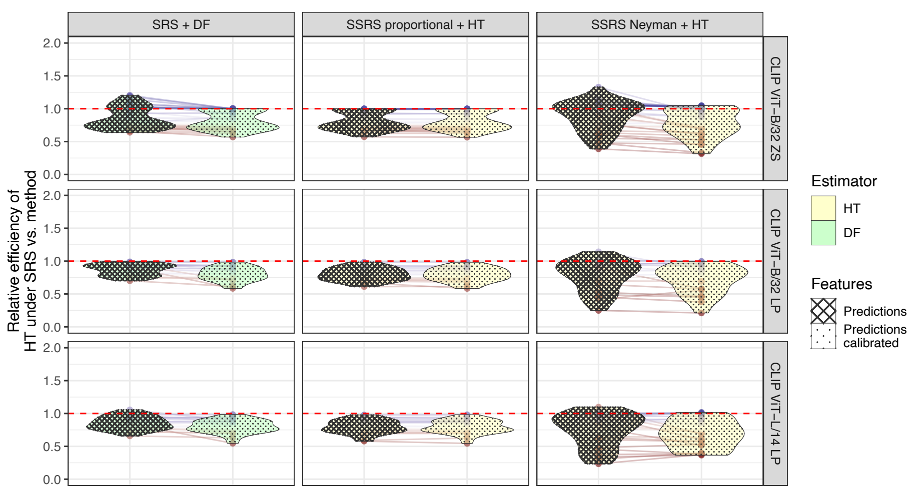

Fig. 5: Comparison of efficiency across sampling designs, estimators, and CLIP models in the zero-shot setting (ZS) and with linear probing (LP). In this figure, we present the results specifically for the proxy $\widehat{Z}$ of Z built on the model being evaluated. For a more detailed explanation of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

8.4 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 4 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.

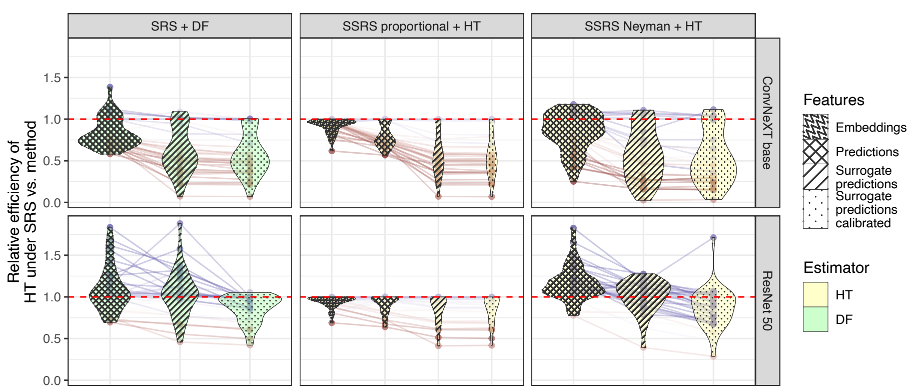

Fig. 6: Comparison of efficiency across sampling designs, estimators, and models. We evaluate the performance of ResNet 50 on the LAION CLIP benchmark tasks using surrogate predictions from ResNet 101. Both models are pretrained on the same data. Similarly, for ConvNeXT, we assess the accuracy of the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

8.5 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.

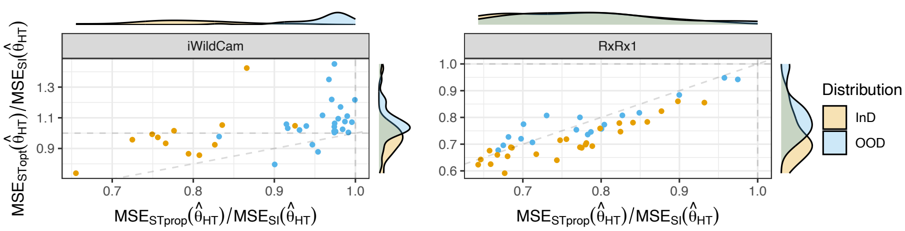

Fig. 7: Comparison of the efficiency of sampling designs and estimators on in-distribution versus out-of-distribution data. The relative efficiencies of `HT` under `SRS` vs. the `HT` estimator under `SSRS` with proportional allocation (horizontal axis) and Neyman allocation (vertical axis) are shown 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`.