Enhancing Cost Efficiency in Active Learning with Candidate Set Query

Anonymous authors
Paper under double-blind review

Abstract

This paper introduces a cost-efficient active learning (AL) framework for classification, featuring a novel query design called *candidate set query*. Unlike traditional AL queries requiring the oracle to examine all possible classes, our method narrows down the set of candidate classes likely to include the ground-truth class, significantly reducing the search space and labeling cost. Moreover, we leverage conformal prediction to dynamically generate small yet reliable candidate sets, adapting to model enhancement over successive AL rounds. To this end, we introduce an acquisition function designed to prioritize data points that offer high information gain at lower cost. Empirical evaluations on CIFAR-10, CIFAR-100, and ImageNet64x64 demonstrate the effectiveness and scalability of our framework. Notably, it reduces labeling cost by 48% on ImageNet64x64.

1 Introduction

Deep neural networks owe much of their success to large-scale annotated datasets (Deng et al., 2009b; Kirillov et al., 2023; OpenAI, 2023; Radford et al., 2021). Scaling datasets is crucial for improving both of their performance (Hestness et al., 2017; Zhai et al., 2022) and robustness (Fang et al., 2022). However, the resources demanded for manual annotation pose a significant bottleneck, particularly in fields requiring expert input like medical data. In response to these challenges, cost-efficient methods for dataset collection, such as semi-automatic labeling (Kim et al., 2024; Qu et al., 2024; Wang et al., 2024), synthetic data generation (Liu et al., 2019a; Tran et al., 2019), and active learning (AL) (Ash et al., 2020; Kirsch et al., 2019; Sener & Savarese, 2018; Settles, 2009; Sinha et al., 2019; Wang & Ye, 2015) have been studied.

This paper investigates AL for classification, where a training algorithm selects informative samples from the data pool and queries annotators for their class labels within a limited budget. We focus on improving the design of annotation queries, emphasizing their critical role. To be specific, we consider image classification of L classes. In the conventional query design, an annotator is asked to choose a class from a list of L classes. Here, the effort needed to review the entire class list and identify the correct class increases as the list size L increases; according to an information-theoretic analysis (Hu et al., 2020), the cost of choosing among L options is $\log_2 L$. To address this issue of growing annotation cost, recent studies (Hu et al., 2020; Kim et al., 2024) employ a 1-bit query design asking annotators to check if the top-1 model prediction is correct. While this simplifies and speeds up annotation, it produces weak supervision incompatible with standard classification loss functions, necessitating specialized losses and algorithms like contrastive loss and semi-supervised learning techniques.

We propose *candidate set query* (CSQ), a novel AL query design that remains cost-efficient with increasing classes and integrates seamlessly with existing loss functions. CSQ presents the annotator with an image and a narrowed set of candidate classes, which is likely to include the ground-truth class. The annotator first searches this small candidate set for the ground-truth class and proceeds to the remaining classes *only if* the ground-truth class is not found during the first search. This query approach can reduce labeling cost by reducing the search space required for annotation, which is particularly effective in scenarios with a wide range of classes where the search space for the annotator could be extensive. Figure 1(*left*) compares CSQ with the conventional query in AL for classification to show its efficiency.

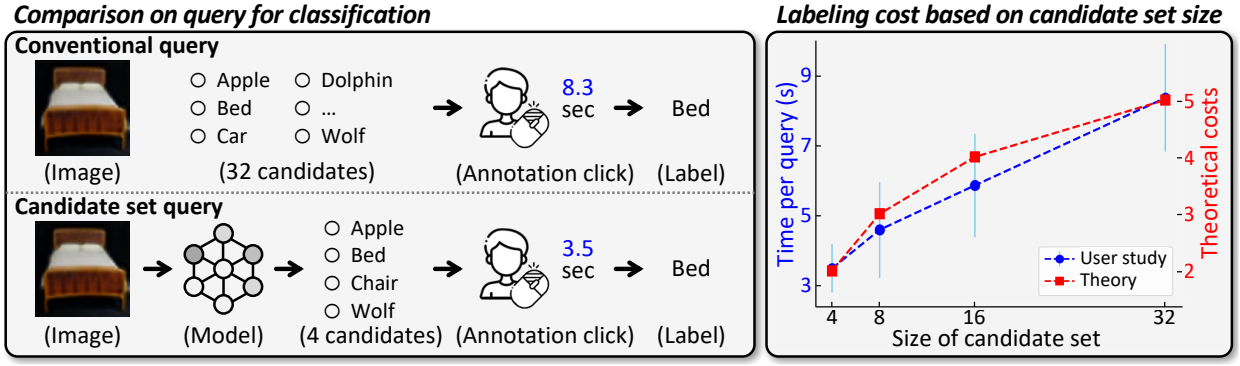


Figure 1: Conventional query versus candidate set query. (*left*) While the conventional query presents all possible options to annotators, CSQ leverages the knowledge of the model to offer narrowed options that are likely to include the ground-truth label, thereby reducing the annotation time. (*right*) By conducting a user study on 40 participants, we demonstrate that the labeling cost increases logarithmically to the candidate set size, which closely aligns with the information-theoretic cost suggested by Hu et al. (2020) with a correlation coefficient of 0.97. Note that as the labeling cost increases per sample, the overall labeling cost increases significantly when multiplied by the total number of labeled samples. Further details of the user study are provided in Sec. 4.2 and Appendix A.

In the CSQ framework, the design of the candidate set is crucial for its effectiveness. On one hand, too many candidates unnecessarily increase the labeling costs. On the other hand, too few candidates are likely to omit the ground-truth class, requiring an additional query to identify the ground-truth class among the remaining classes, which is more expensive than the conventional query. To enhance the effectiveness of the CSQ framework, we propose to construct candidate sets guided by prediction uncertainty from a trained model using conformal prediction (Shafer & Vovk, 2008; Angelopoulos et al., 2023). Conformal prediction aims at constructing a set of predictions including the true class, where each set is properly sized based on the certainty of the model about the input. This strategy enables flexible adjustment of the candidate set for each sample, expanding it for an uncertain sample to include the true label and shrinking it for a more certain one to reduce the labeling cost. Furthermore, we optimize the level of certainty in conformal prediction to minimize the labeling cost for each round. Therefore, this candidate set construction adapts to the increasing accuracy of the model over successive AL rounds, refining the candidate set as the model improves.

Last but not least, we propose a new acquisition function designed to maximize the cost efficiency of CSQ. Conventional acquisition functions in AL are designed to favor samples with high estimated information gain, assuming uniform annotation costs across all samples. On the other hand, in CSQ, the labeling cost for each sample varies according to the size of its candidate set. Thus, we propose an acquisition function that evaluates samples based on the ratio of estimated information gain to labeling cost. Specifically, we combine the conventional acquisition function score, which indicates the estimated information gain, with the estimated cost derived from the candidate set, favoring samples that maximize information gain per unit cost. This cost-efficient acquisition function can incorporate with any sample-wise acquisition score, ensuring the selection of both informative and cost-efficient samples.

The proposed method achieves state-of-the-art performance on CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), and ImageNet64x64 (Chrabaszcz et al., 2017). We verify the effectiveness and generalizability of CSQ through extensive experiments with varying datasets, acquisition functions, and budgets. Notably, CSQ achieves the same performance as the conventional query on ImageNet64x64 at only 48% of the cost, showing its scalability. Our ablation studies demonstrate that both our candidate set construction and sampling strategy contribute to the performance. Furthermore, the necessity of CSQ is demonstrated by a user study involving 40 participants. In short, the main contribution of this paper is four-fold:

- We propose a novel query design for active learning, where the annotator is presented with an image and a narrowed set of candidate classes that is likely to include the ground-truth class. This approach, termed CSQ, reduces labeling cost by minimizing the search space the annotator needs to explore.
- To maximize the advantage of CSQ, we propose to utilize conformal prediction to dynamically generate small yet reliable candidate sets optimized to reduce labeling costs, adapting to the evolving model throughout successive AL rounds.
- We propose a new acquisition function that prioritizes a data point expected to have high information gain relative to its labeling cost, enhancing cost efficiency.
- The proposed framework achieved state-of-the-art performance on diverse image classification datasets, CIFAR-10, CIFAR-100, and ImageNet64x64, showing its effectiveness and generalizability.

2 Related Work

Acquisition functions in AL. The key to AL is to select and annotate the most informative samples Settles (2009); Dasgupta (2011); Hanneke et al. (2014). To assess informativeness, various acquisition functions have been proposed, considering either the uncertainty of model predictions (Asghar et al., 2017; He et al., 2019; Ostapuk et al., 2019; Fuchsgruber et al., 2024; Kim et al., 2024; Cho et al., 2024; Kim et al., 2023; Wang et al., 2022), diversity in feature space (Sener & Savarese, 2018; Sinha et al., 2019; Yehuda et al., 2022), or both (Ash et al., 2020; Hwang et al., 2022; Wang & Ye, 2015; Wang et al., 2019; Hacohen et al., 2022; Hacohen & Weinshall, 2023;?). Disagreement-based AL and its variants are supported by rigorous theoretical learning guarantees (Hanneke et al., 2014; Krishnamurthy et al., 2019). However, these methods assume uniform sample costs and select based solely on the amount of information. We emphasize that the labeling cost required for each sample varies and prioritize samples offering the best information-to-cost ratio.

Realistic cost model for AL. While we follow Hu et al. (2020) in modeling annotation cost based on the size of the labeling options using an information-theoretic perspective, other works have explored additional factors influencing real-world annotation cost, such as annotator behavior and interaction complexity (Settles et al., 2008; Arora et al., 2009; Wallace et al., 2011; Herde et al., 2021) For example, Settles et al. (2008) and Arora et al. (2009) show that instance, task, and user features jointly affect annotation time, while Wallace et al. (2011) propose a cost-saving strategy where less experienced annotators can pass uncertain cases to more experienced ones. Herde et al. (2021) provide a taxonomy of cost-aware active learning approaches under more realistic assumptions. These studies highlight that annotation cost depends on multiple factors and varies across settings. While our size-based proxy is effective and supported by user study, richer cost models may better capture real-world annotation behavior and further improve performance.

Conformal prediction (CP). CP enables us to quantify uncertainty in predictions with an associated confidence level (Shafer & Vovk, 2008). Recent advances in CP empower classifiers to generate predictive sets that include the ground-truth label with a probability chosen by the user (Angelopoulos et al., 2020; 2023). In the field of AL, nonconformity measurements from CP are employed in the acquisition function to select informative samples (Matiz & Barner, 2020). In contrast, we utilize CP not only to develop a cost-efficient acquisition function but also to design an efficient candidate set query reducing the labeling cost.

3 Proposed Method

We consider general classification tasks such that for input $\mathbf{x}$ and a categorical variable $y \in \mathcal{Y} = \{1, 2, \ldots, L\}$, a model parameterized by $\boldsymbol{\theta}$ predicts the class of the input as $\arg\max_{y \in \mathcal{Y}} P_{\boldsymbol{\theta}}(y|\mathbf{x})$. We study an active learning (AL) scenario conducted over R rounds. In each round r, a budget of B samples with high acquisition function values is actively selected from the unlabeled data pool $\mathcal{X}$. This actively selected set $\mathcal{A}_r$ is then labeled by an annotator to form the labeled dataset $\mathcal{D}_r$ with labeling cost C_r, and is used to update the model. Let $\boldsymbol{\theta}_r$ denote the model trained on the accumulated labeled data up to round r, $\bigcup_{i=0}^{r} \mathcal{D}_i$. Our goal

Algorithm 1 Cost-efficient active learning with candidate set query

Require: The number of AL rounds R, per-round budget B, unlabeled data pool $\mathcal{X}$, initial labeled dataset $\mathcal{D}_0$.
1: Train the initial model $\boldsymbol{\theta}_0$ on $\mathcal{D}_0$.
2: **for** $r = 1, 2, \ldots, R$ **do**
3: Estimate sample-wise labeling cost of $\mathbf{x} \in \mathcal{X}$.
4: Select B samples $\mathcal{A}_r \subset \mathcal{X}$ considering both their estimated labeling cost and informativeness. ▷ Sec. 3.3
5: Construct candidate set $\hat{Y}(\mathbf{x})$ for each $\mathbf{x} \in \mathcal{X}$. ▷ Sec. 3.2
6: Query annotator for label y of $\mathbf{x} \in \mathcal{A}_r$ using candidate set $\hat{Y}(\mathbf{x})$ to form $\mathcal{D}_r$.
7: Get model $\boldsymbol{\theta}_r$ trained on $\bigcup_{i=0}^{r} \mathcal{D}_i$.
8: **end for**
9: **Return** Final model $\boldsymbol{\theta}_R$.

is to maximize the performance of $\boldsymbol{\theta}_r$, while minimizing the accumulated cost $\bigcup_{i=0}^{r} C_i$. The key aspect of the proposed method is candidate set query (CSQ), which reduces C_r by narrowing the set of candidate classes presented to annotators. For simplicity, we omit the round index r from $\boldsymbol{\theta}_r$ in the remainder of this section.

In the following, we first introduce CSQ and discuss its efficiency in labeling cost (Sec. 3.1). Then, we present a method to construct a candidate set based on the prediction uncertainty of a trained model for a given sample (Sec. 3.2). Lastly, we introduce an acquisition function designed to consider cost efficiency as well as information gain (Sec. 3.3). The overall pipeline of the CSQ framework combined with our cost-efficient sampling is summarized in Algorithm 1.

3.1 Candidate set query

CSQ for an instance $\mathbf{x}$ is associated with a (non-empty) candidate set $Y(\mathbf{x}) \subseteq \mathcal{Y}$ such that $1 \leq |Y(\mathbf{x})| \leq L$. CSQ first asks the annotator to choose the ground-truth class in $Y(\mathbf{x})$ if it exists, or to verify the absence of the ground-truth label in $Y(\mathbf{x})$, *i.e.*, the annotator is first asked to pick an option out of $(k+1)$ choices, where $k = |Y(\mathbf{x})|$. Only if the absence of the ground-truth class in the candidate set is verified, the annotator is further asked to select the ground-truth class from the remaining ones $\mathcal{Y} \setminus Y(\mathbf{x})$. In short, CSQ asks the following query with an image to the oracle.

Select the ground-truth class from the candidate set, or choose "None of the above."
(only if "None of the above" *is chosen): Select the ground-truth from the remaining classes.*

Following the information-theoretic cost model of (Hu et al., 2020) and the user-study results in Table 1, we model the cost of selecting one label from k candidates as $\log_2 k$. Then, the labeling cost $\Gamma(Y(\mathbf{x}), y)$ of CSQ for input $\mathbf{x}$, ground-truth label y, and candidate set $Y(\mathbf{x}) \subseteq \mathcal{Y}$ can be obtained as:

$$\Gamma(Y(\mathbf{x}), y) = \begin{cases} \log_2(k+1) & \text{if } y \in Y(\mathbf{x}) \\ \log_2(k+1) + \log_2(L-k) & \text{otherwise} \end{cases}. \tag{1}$$

The conventional query in AL is a special case of CSQ where $Y(\mathbf{x}) = \mathcal{Y}$, and it is inefficient since the annotator must search through the entire set of size L with a cost of $\log_2 L$. The following theorem reveals the condition under which the expected cost of CSQ offers an improvement over that of the conventional query.

Theorem 3.1. *Assume the information-theoretic cost model (Hu et al., 2020) of selecting one out of L possible options to be $\log_2 L$. Let $L \geq 2$ be the number of classes, $k = |Y(\mathbf{x})|$, and α be the probability that the candidate set $Y(\mathbf{x})$ does not include the ground-truth class of instance $\mathbf{x}$. Denote by C_{con} and C_{csq} the expected costs of the conventional query and the candidate set query, respectively. If*

$$\frac{\log_2(k+1)}{\log_2 L} < 1 - \alpha\,, \tag{2}$$

then the candidate set query is strictly cheaper, i.e., $C_{\mathrm{csq}}(L, \mathbf{x}, \alpha) < C_{\mathrm{con}}(L, \mathbf{x})$.

Proof. Recalling the definition of α, we have $C_{\text{csq}}(L, \mathbf{x}, \alpha) = (1-\alpha)\log_2(k+1) + \alpha\{\log_2(k+1) + \log_2(L-k)\}$ from Eq. (1). As $L - k < L$, the cost ratio of $C_{\text{csq}}(L, \mathbf{x}, \alpha)$ to $C_{\text{con}}(L, \mathbf{x})$ for instance $\mathbf{x}$ is induced as:

$$\begin{aligned} \frac{C_{\text{csq}}(L, \mathbf{x}, \alpha)}{C_{\text{con}}(L, \mathbf{x})} &= \frac{\log_2(k+1) + \alpha \log_2(L-k)}{\log_2 L} \\ &< \frac{\log_2(k+1)}{\log_2 L} + \alpha \, . \end{aligned} \tag{3}$$

Although we adopt the cost model from Hu et al. (2020), Theorem 3.1 holds for any cost model that increases monotonically with the number of options.

Remark 3.2. *If we constrain all candidate set sizes k to be fixed, then $1 - \alpha$ corresponds to the top-k accuracy p_k of the model. Therefore, when $p_k \geq \log_L(k+1)$, CSQ consistently offers an improvement over the conventional query in the expected labeling cost. For example, in datasets such as CIFAR-10 ($L = 10$), CIFAR-100 ($L = 100$), and ImageNet ($L = 1000$), if the model has a top-1 accuracy (*i.e., *$k = 1$) of at least 30.1%, 15.1%, and 10.0% respectively, then CSQ always provides an improvement.*

The above proof and remark demonstrate that under moderate conditions, CSQ is more efficient than the conventional query. As described in Eq. (3), the cost of CSQ decreases as both α and k become smaller. However, since k and α are inversely related, balancing the trade-off between α and k is essential to fully leverage CSQ. Also, fixing candidate set sizes as in Remark 3.2 is suboptimal because it does not consider the difficulty of individual samples. In the following section, we introduce our candidate set construction method, which both reflects the uncertainty of each sample and automatically balances the trade-off between α and k.

3.2 Construction of cost-efficient candidate set

As shown in Eq. (1) and Theorem 3.1, a candidate set needs to be both small and accurate in covering the ground-truth class. To do so, we propose using conformal prediction (Romano et al., 2020) to get a reliable and cost-optimized candidate set using the trained model $\boldsymbol{\theta}$ of the previous round.

Calibration set collection. Conformal prediction requires a labeled set for calibration that has not been used during the model training phase; this set must follow the same distribution as the target dataset for prediction (Vovk et al., 1999; Angelopoulos et al., 2023). To achieve this, we randomly select n_{cal} samples from the actively selected data $\mathcal{A}_r$ and annotate them within the given budget to form $\mathcal{D}_{\text{cal}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_{\text{cal}}}$. The calibration set $\mathcal{D}_{\text{cal}}$ is used for conformal prediction and candidate set optimization, which will be explained in the following sections. Note that $\mathcal{D}_{\text{cal}}$ also contributes to model training after the candidate set construction.

Conformal prediction. Using the model $\boldsymbol{\theta}$ from the previous round and calibration set $\mathcal{D}_{\text{cal}} \subset \mathcal{A}_r$ of size n_{cal}, we define a collection of conformal scores $\mathbf{s} := (s_i)_{i=1}^{n_{\text{cal}}}$, where $s_i := 1 - P_{\boldsymbol{\theta}}\big(y_i \mid \mathbf{x}_i\big)$ for $(\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{cal}}$. Then, we obtain the $(1-\alpha)$ empirical quantile $\hat{Q}(\alpha)$ of $\mathbf{s}$, indicating that at least $100 \times (1-\alpha)\%$ of the scores in $\mathbf{s}$ are smaller than $\hat{Q}(\alpha)$. This empirical quantile $\hat{Q}(\alpha)$ is given as,

$$\hat{Q}(\alpha) := \min_{s \in \mathbf{s}} \left\{ s : \frac{1}{n_{\text{cal}}} \sum_{s' \in \mathbf{s}} \big(\mathbb{1}[s' \leq s]\big) \geq 1 - \alpha \right\} \, , \tag{4}$$

where $\alpha \in (0, 1)$ is an error rate hyperparameter, and $\mathbb{1}[\cdot]$ is an indicator function. Then, we define the candidate set $\hat{Y}_{\boldsymbol{\theta}}(\mathbf{x}, \alpha)$ for unlabeled data $\mathbf{x}$ using conformal prediction as follows:

$$\hat{Y}_{\boldsymbol{\theta}}(\mathbf{x}, \alpha) := \big\{ y : P_{\boldsymbol{\theta}}(y|\mathbf{x}) \geq 1 - \hat{Q}(\alpha), \; y \in \mathcal{Y} \big\} \, . \tag{5}$$

Previous study (Vovk et al., 1999; Angelopoulos et al., 2023) proved that the candidate set includes the true label with the probability not less than $1 - \alpha$, which is,

$$P\big(y \in \hat{Y}_{\boldsymbol{\theta}}(\mathbf{x}, \alpha)\big) \geq 1 - \alpha \, . \tag{6}$$

This ensures the inclusion of the ground-truth classes even under model overconfidence, while adaptively reflecting uncertainties throughout the AL process. Without loss of generality, we consider $\hat{Y}_{\boldsymbol{\theta}}(\mathbf{x}, 0) = \mathcal{Y}$, where $\mathcal{Y}$ corresponds to the conventional query. More detailed procedure of conformal prediction is in Appendix C.

Cost-optimized error rate selection. The proposed candidate set construction method (Eq. (5)) adapts the size of the candidate set for each sample based on its predicted uncertainty. While this allows the candidate set to be both compact and reliable, it requires manually adjusting the hyperparameter α. To eliminate the need for manual tuning, we introduce an automatic selection scheme that optimizes α at each AL round by minimizing the expected labeling cost on the calibration set, which serves as a pseudo-validation set. To be specific, α is optimized by

$$\alpha^* := \underset{\alpha \in [0,1)}{\arg\min} \sum_{(\mathbf{x},y) \in \mathcal{D}_{\text{cal}}} \Gamma(\hat{Y}_{\boldsymbol{\theta}}(\mathbf{x}, \alpha), y) \ , \tag{7}$$

where $\Gamma(\cdot, y)$ is the labeling cost in Eq. (1). We implement this optimization using a grid search over predefined options of α. The optimization is computationally efficient, requiring only negligible computation as shown in Figure 7. This approach not only eliminates the reliance on hand-tuned hyperparameters but also helps construct candidate sets in a more cost-efficient manner, as the selected α^* is tailored to minimize the labeling cost under the proposed cost model for each round. Since the optimization in Eq. (7) naturally considers the conventional query as a special case of CSQ at $\alpha = 0$, CSQ is at least as efficient as, and in general more efficient than, the conventional query.

Note that to construct the candidate set query, the calibration set $\mathcal{D}_{\text{cal}}$ is required to calculate $(1 - \alpha^*)$ quantile in Eq. (4). Thus, when getting annotations of $\mathcal{D}_{\text{cal}}$ in the calibration set collection step, the candidate set query of the current round cannot be applied. To avoid this circular dependency, the quantile from the previous round is used when labeling $\mathcal{D}_{\text{cal}}$.

3.3 Cost-efficient acquisition function

Since the labeling cost of each sample varies in CSQ, we propose to consider the cost for annotation in the acquisition function. We implement an acquisition function that evaluates samples based on the ratio of the estimated information gain to the estimated labeling cost. The information gain is quantified using an established acquisition score, entropy, BADGE (Ash et al., 2020), or SAAL (Kim et al., 2023), though our approach is compatible with other acquisition scores as well, not just these. Given a conventional acquisition score $g_{\text{score}}(\mathbf{x})$, the proposed cost-efficient acquisition function g_{cost} is given by,

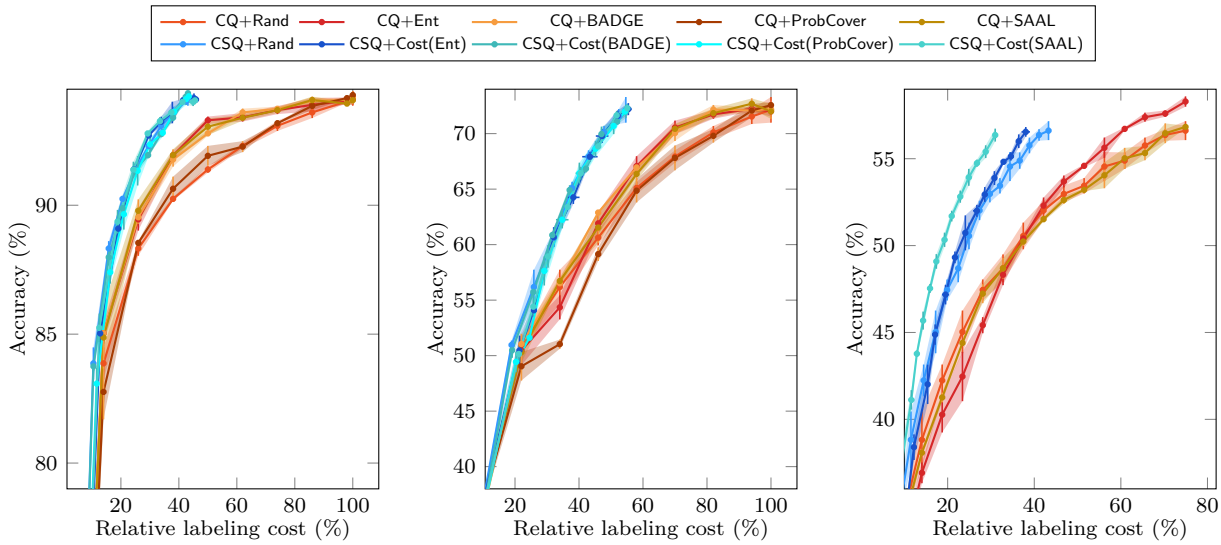
$$g_{\text{cost}}(\mathbf{x}) := \frac{(1 + g_{\text{score}}(\mathbf{x}))^d}{\log_2(k+1) + \alpha^* \log_2(L-k)} \ , \tag{8}$$

where d is a hyperparameter adjusting the influence of $g_{\text{score}}(\mathbf{x})$ and α^* is the optimized error rate hyperparameter obtained by Eq. (7). The denominator is an expected cost derived from our cost model (Eq. (1)), considering two cases: the correct label is included or excluded from the candidate set, which is $(1 - \alpha^*) \log_2(k+1) + \alpha^* \{\log_2(k+1) + \log_2(L-k)\}$. This expected cost assumes the candidate set to include the ground-truth class with a probability of $1 - \alpha^*$, which is supported by the coverage guarantee in Eq. (6).

4 Experiments

4.1 Experimental setup

Datasets. We use three image classification datasets: CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), and ImageNet64x64 (Chrabaszcz et al., 2017). CIFAR-10 comprises 50K training and 10K validation images across 10 classes. CIFAR-100 contains the same number of images as CIFAR-10, but is associated with 100 classes. ImageNet64x64 is a downsampled version of ImageNet (Deng et al.,



(a) CIFAR-10 (b) CIFAR-100 (c) ImageNet64x64

Figure 2: Accuracy (%) versus relative labeling cost (%) for conventional query (CQ) and the proposed candidate set query (CSQ) with different acquisition functions. CSQ approaches (blue lines) consistently outperform the CQ baselines (red lines) by a significant margin across various cost budgets, acquisition functions, and datasets[1].

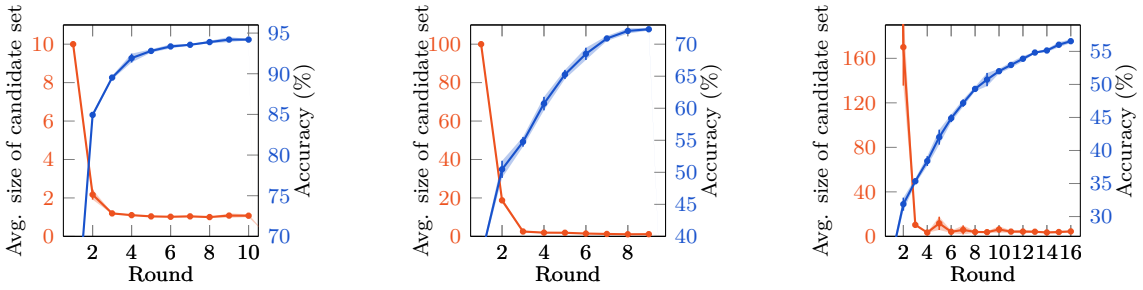
2009a) with a resolution of 64×64, which consists of 1.2M training and 50K validation images with 1000 classes. Following previous studies, we evaluate a model using the validation split of each dataset.

Implementation details. For CIFAR-10 and CIFAR-100, we adopt ResNet-18 (He et al., 2016) as a classification model. We train it for 200 epochs using AdamW (Loshchilov & Hutter, 2019) optimizer with an initial learning rate of 1e−3, decreasing by a factor of 0.2 at epochs 60, 120, and 160. We apply a weight decay of 5e−4 and a data augmentation consisting of random crop, random horizontal flip, and random rotation. For ImageNet64x64, we adopt WRN-36-5 (Zagoruyko, 2016), and train it for 30 epochs using AdamW optimizer with an initial learning rate of 8e−3. We apply a learning rate warm-up for 10 epochs from 2e−3. After the warm-up, we decay the learning rate by a factor of 0.2 every 10 epochs. We adopt random horizontal flip and random translation as data augmentation. For all the datasets, we use Mix-up (Zhang et al., 2018), where a mixing ratio is sampled from $\text{Beta}(1, 1)$. We set the size of the calibration dataset n_{cal} to 500 for CIFAR-10 and CIFAR-100, and 5K for ImageNet64x64. For all datasets and acquisition functions, hyperparameter d in Eq. (8) is set to 0.3.

Active learning protocol. For CIFAR-10, we conduct 10 AL rounds of consecutive data sampling and model updates, while for CIFAR-100, we perform 9 AL rounds. In both cases, the per-round budget is 6K images. For ImageNet64x64, we conduct 16 AL rounds with a per-round budget of 60K images. The detailed budget configuration for the three datasets is shown in Table 5. In the initial round, we randomly sample 1K images for CIFAR-10, 5K images for CIFAR-100, and 60K images for ImageNet64x64. In each round, the model is evaluated based on two factors: its accuracy (%) on the validation set, and the accumulated annotation cost required to train it. The annotation cost is defined as a relative labeling cost (%) compared to the cost of labeling the entire training set using the conventional query, given by $N \log_2 L$, where N is the size of the entire training set, and L is the number of classes. We conduct all experiments with three independent trials with different random seeds and report the mean and standard deviation to ensure reproducibility.

Baseline methods. We compare our candidate set query (CSQ) with the conventional query (CQ) in combination with various sampling strategies. To be specific, we employ random (Rand), entropy (Ent), BADGE (Ash et al., 2020), ProbCover (Yehuda et al., 2022), and SAAL (Kim et al., 2023) as the sampling

[1]Unfortunately, for ImageNet64x64, we exclude the BADGE and ProbCover acquisition baselines due to their computational intractability.



(a) CIFAR-10 (b) CIFAR-100 (c) ImageNet64x64

Figure 3: Average size of the candidate set and accuracy (%) of our method with cost-efficient entropy sampling in varying rounds on CIFAR-10, CIFAR-100, and ImageNet64x64. Note that the candidate set includes all classes at the initial round, but our candidate set design effectively reduces its size over successive active learning rounds, adapting dynamically as the model accuracy improves.

strategies. Cost($\cdot$) indicates the proposed cost-efficient sampling (Eq. (8)) using conventional acquisition scores; *e.g.*, Cost(SAAL) is the one combined with SAAL. We denote the combination of the query and sampling method with '+', *e.g.*, CSQ+Rand is a candidate set query combined with random sampling.

4.2 Experimental results

Candidate set query vs. Conventional query. In Figure 2, we compare the performance of the candidate set query (CSQ) with the conventional query (CQ) on CIFAR-10, CIFAR-100, and ImageNet64x64 with different acquisition functions. CSQ approaches consistently outperform the CQ approaches across various acquisition functions and datasets, demonstrating the general effectiveness of our method. Notably, CSQ reduces the labeling cost by 43%, 54%, and 48% on CIFAR-10, CIFAR-100, and ImageNet64x64, compared to CQ, respectively. This is promising as it shows that the same volume of labeled data can be obtained at roughly half the cost, without introducing any label noise or sample bias. Notably, the performance gain of CSQ increases as the model improves, as it is tailored to the improved model. In the appendix, we also present experiments on a text classification task (Figure 13) showing the generalization ability of the proposed method to the natural language domain. Additionally, we provide the zoomed version of Figure 2 in Figure 16 and Figure 17.

Progressive reduction in candidate set size. The effectiveness of CSQ stems from its ability to reduce labeling costs through smaller candidate sets. To demonstrate this, Figure 3 shows the average size of the candidate sets and accuracy (%) of our method with varying AL rounds on CIFAR-10, CIFAR-100, and ImageNet64x64. After the first round, CSQ achieves a sufficiently small candidate set size and continues to reduce it as accuracy improves, thereby enhancing labeling efficiency.

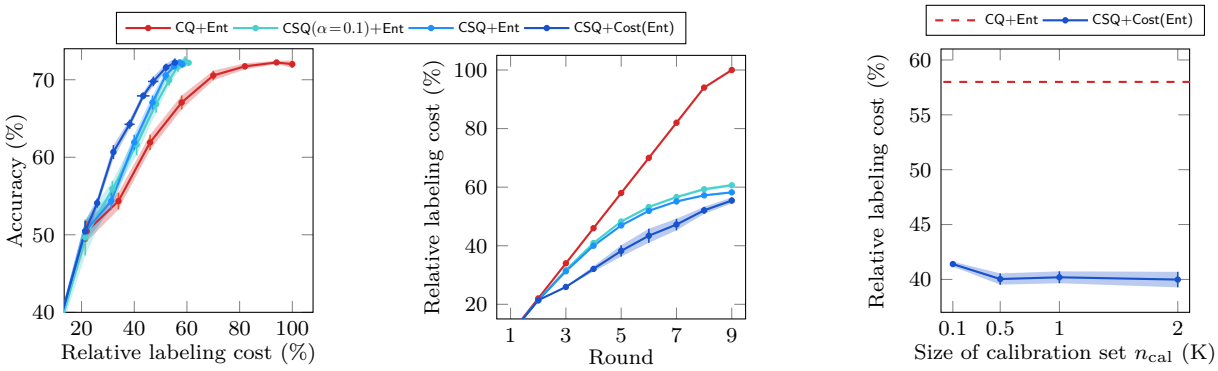
Empirical validation for our cost model. We conduct a user study with 40 annotators who label samples using candidate sets of various sizes; see Appendix A for more details. The results in Table 1 suggest that shrinking candidate sets improves both labeling efficiency and accuracy. They also align closely with the theoretical cost (Hu et al., 2020), as shown in Figure 1(*right*).

4.3 Ablation studies

Contribution of each component. Figure 4a demonstrates the contribution of each component in our method across varying AL rounds: candidate set query (Eq. (5)), cost reduction from α^* (Eq. (7)), and the proposed acquisition function (Eq. (8)). The results show consistent performance improvements from each component in every round. The performance gap between CQ+Ent and CSQ($\alpha=0.1$)+Ent verifies the efficacy of proposed CSQ framework, which provides the largest improvement. The gap between CSQ($\alpha=0.1$)+Ent and CSQ+Ent shows the impact of α optimization, offering modest but steady gains across rounds. Finally, the gap between CSQ+Ent and CSQ+Cost(Ent) shows the effectiveness of our acquisition function, particularly from 4 to 6 rounds.

Table 1: User-study results on a fixed image set with different numbers of class options presented to the annotator. A smaller option set yields both faster annotation and higher accuracy. The empirical trends also align closely with the theoretical cost curves in Figure 1(*right*). In all experiments, we treat annotation cost as proportional to annotation time, which we consider a more practical measure than simply counting labeled examples.

Class options (#)	4	8	16	32
Annotation Time (s)	**69.4**$_{\pm 13.8}$	91.5$_{\pm 27.3}$	116.9$_{\pm 29.6}$	166.9$_{\pm 30.8}$
Accuracy (%)	**100.0**$_{\pm 0.0}$	98.5$_{\pm 3.2}$	99.5$_{\pm 1.5}$	95.5$_{\pm 5.2}$



(a) Contribution of each component

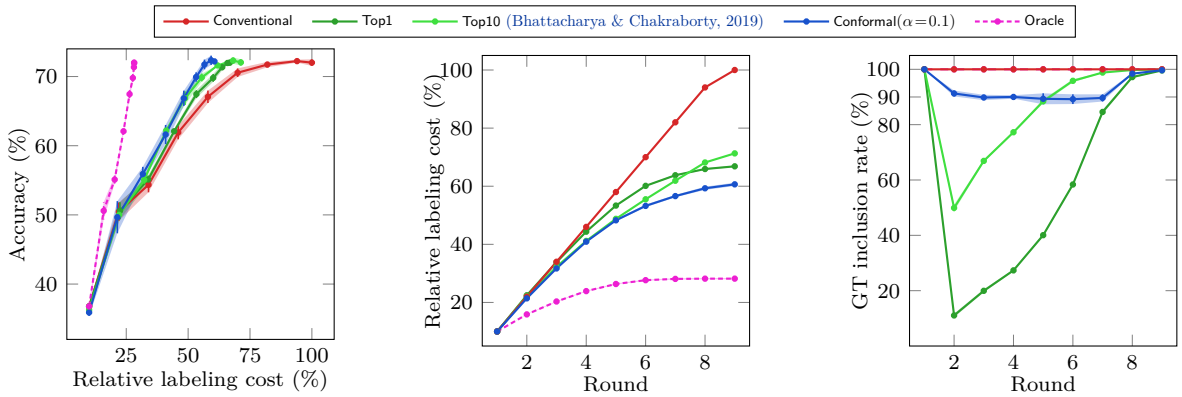
(b) Impact of calibration set size

Figure 4: (a) Contribution of each component of our method, measured by accuracy (%) versus relative labeling cost (%) (*left*), and relative labeling cost (%) versus AL rounds (*right*) on CIFAR-100. The results compare the full method (CSQ+Cost(Ent)), the method without acquisition function in Eq. (8) (CSQ+Ent), without α optimization in Eq. (7), where α is fixed to 0.1 (CSQ($\alpha=0.1$)+Ent), and without CSQ (CQ+Ent). All components of our method lead to steady performance improvement over varying rounds. (b) Relative labeling cost (%) at the fifth round with varying calibration set sizes n_{cal} in Eq. (4) on CIFAR-100. The dashed line indicates the relative labeling cost (%) of CQ+Ent. Our method shows consistent performance with varying calibration set sizes.

Impact of calibration set size. In Figure 4b, we evaluate the relative labeling cost (%) at the fifth round with varying calibration set sizes n_{cal} in Eq. (4) to assess its impact on the performance on CIFAR-100. As shown in Figure 4b, our method shows consistent performance, varying by less than 2%p as the calibration set size changes from 0.1K to 2K, and significantly outperforms the baseline.

Detailed ablation study on candidate set design. Figure 5 illustrates the effectiveness of using conformal prediction (Conformal ($\alpha=0.1$)) for candidate set construction on CIFAR-100, compared to baselines: Conventional (using all classes), Top1 (top-1 prediction), Top10 (top-10 predictions), and Oracle (smallest top-k set always containing the ground truth). Note that Oracle represents an unattainable upper bound requiring knowledge of the ground truth. Top10 is a variant of the n-ary query (Bhattacharya & Chakraborty, 2019) baseline. For consistency, we fixed $\alpha = 0.1$ in Eq. (5). Figures 5a and 5b show that conformal prediction consistently reduces labeling cost compared to the baselines. While Top10 is effective in the early rounds and Top1 becomes more efficient as the model improves, our method adapts and outperforms all baselines in every round. Figure 5c demonstrates that with $\alpha = 0.1$, our method includes the ground-truth class in over 90% of cases, aligning with Eq. (6), while the top-k baselines show lower inclusion rates, especially in early and middle rounds. This demonstrates that conformal prediction effectively adjusts candidate set sizes based on sample uncertainty, ensuring ground-truth inclusion and improving labeling efficiency.

Detailed ablation study on cost-efficient acquisition function. In Table 2, we investigate the impact of the proposed cost-efficient sampling (Sec. 3.3) on CIFAR-100, where performance is measured by accuracy per cost (*i.e.*, accuracy divided by relative labeling cost). This metric reflects how efficiently a method achieves high accuracy under a fixed annotation budget—higher values indicate better cost-effectiveness.



(a) Accuracy (b) Relative labeling cost (c) GT inclusion rate

Figure 5: Impact of the candidate set design evaluated on CIFAR-100 using the conventional query with all classes (Conventional), top-1 prediction from model (Top1), top-10 predictions from model (Top10), our method with conformal prediction with fixed $\alpha = 0.1$ (Conformal($\alpha = 0.1$)), and the smallest top-k prediction sets always including the ground-truth class (Oracle). Note that Top10 is a variant of the n-ary query (Bhattacharya & Chakraborty, 2019) baseline. For comparison, the same entropy sampling is used to keep the accuracy at each round constant, focusing solely on the labeling cost and isolating the effect of the candidate set design. (a) Our method constantly outperforms the baselines both in accuracy (%) and relative to labeling cost (%). (b) Our design achieves a greater reduction in labeling cost compared to the baselines. (c) Our candidate set effectively includes the ground-truth class in over 90% of cases ($= 1 - \alpha$), even when model accuracy is low.

Table 2: Effectiveness of the proposed cost-efficient sampling (Cost($\cdot$)) on CIFAR-100, reported as *accuracy per cost*. All methods employ the same candidate set query (CSQ) framework, and we simply replace the acquisition function baselines with their Cost($\cdot$) variants. Our sampling consistently improves the performance across various acquisition functions and AL rounds. The best results are highlighted in bold.

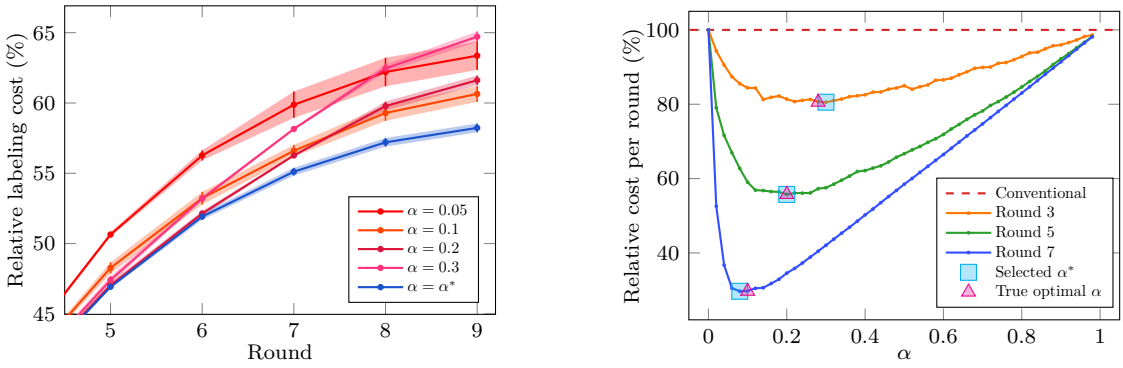
Acquisition	2nd round	3rd round	4th round	5th round	6th round	7th round	8th round	9th round
Ent	2.36	1.74	1.55	1.43	1.36	1.30	1.26	1.24
Cost(Ent)	2.36	**2.09**	**1.89**	**1.68**	**1.56**	**1.48**	**1.37**	**1.30**
BADGE	2.51	1.94	1.69	1.51	1.40	1.34	1.29	1.27
Cost(BADGE)	**2.64**	**2.17**	**1.92**	**1.75**	**1.58**	**1.49**	**1.37**	**1.32**
ProbCover	2.43	1.72	1.60	1.55	1.47	1.39	1.33	1.30
Cost(ProbCover)	2.43	**2.10**	**1.98**	**1.79**	**1.66**	**1.52**	**1.39**	**1.32**
SAAL	**2.37**	1.83	1.55	1.43	1.37	1.32	1.28	1.25
Cost(SAAL)	2.36	**2.12**	**1.94**	**1.74**	**1.64**	**1.50**	**1.38**	**1.31**

The proposed cost-efficient sampling strategy consistently improves performance across different acquisition functions, with gains observed throughout various AL rounds. A more detailed accuracy versus labeling cost graph is illustrated in Figure 10.

Ablation study on cost-optimized error rate selection. In Figure 6a, we present the impact of cost-optimized error rate selection as in Eq. (7), evaluated on CIFAR-100 using entropy sampling, in terms of relative labeling cost (%). Compared to the baselines using hand-picked error rate values, the cost-optimized error rate $\alpha = \alpha^*$ from the proposed method consistently reduces labeling cost across all AL rounds. This demonstrates that the proposed method reduces the need for manual tuning of the error-rate hyperparameter and instead automatically selects an error rate optimized for each AL round.

4.4 In-depth analysis

Quality of the optimized error rate α^*. In Figure 6b, we compare the optimized error rate (α^*, blue squares) selected by our method with the true optimal error rate (pink triangles), showing their relative



(a) Impact of cost-optimized error rate

(b) Quality of selected error rate α^*

Figure 6: Impact of cost-optimized error rate selection as in Eq. (7), evaluated on CIFAR-100 with entropy sampling. (a) Relative labeling cost (%) versus AL rounds with different error rate α compared with the α^* selected by the proposed method (Eq. (7)). This shows our method removes manual tuning by automatically choosing an adaptive error rate at each round. (b) Relative labeling cost per round (%) versus α across various AL rounds. The cost curves reveal that selecting the right error rate greatly lowers labeling cost, with the benefit growing in later rounds. Pink triangles mark the true optimal α, blue squares mark α^* from our method. The selected α^* remains consistently close to the true optimum, enabling each round to sharply reduce labeling cost.

labeling costs (%) across different AL rounds. As indicated by the labeling cost curves for each error rate, choosing an effective error rate significantly impacts cost reduction, and this impact becomes more prominent at later rounds. Our proposed method adaptively selects an optimized error rate (α^*) close to the true optimal error rate at each AL round, substantially reducing labeling costs.

Analysis on computational complexity. Figure 7 plots the wall-clock time measured in *minutes* of each strategy as the size of unlabeled pool from ImageNet64x64 grows from 10K to 0.7M images. The dashed curve isolates the cost of the candidate set query (CSQ) itself and reveals a flat line: constructing candidate sets is *constant time*, independent of the unlabeled pool size. All entropy-based samplings, including our proposed CSQ+Cost(Ent), exhibit the expected *linear* slope. This confirms that both the candidate set query and cost-efficient sampling add negligible overhead, preserving the linear complexity of standard entropy sampling. By contrast, CQ+BADGE rises steeply owing to the k-means++ seeding inside BADGE[2], even with an accelerated implementation[3].

Table 3 gives an absolute breakdown at the largest pool size (1.21 M images). Although conventional queries (CQ+Ent, CQ+BADGE) have zero query cost—they simply show all classes to the oracle—our CSQ adds only <0.05 minutes, a negligible overhead that confirms the constant-time trend in the Figure 7. Likewise, replacing plain entropy sampling with its cost-efficient variant (Cost(Ent)) increases sampling time by less than 0.01 min, preserving the same linear complexity. Both components of our pipeline are light—CSQ for querying and Cost($\cdot$) for sampling—so the full method (CSQ+Cost(Ent)) remains scalable even on million-scale pools.

Examples of constructed candidate sets. In Figure 8, we present input images and their corresponding candidate sets on ImageNet64x64. Thanks to the conformal prediction, the proposed method allows flexible adjustments of the candidate set for each sample.

5 Conclusion

We have introduced candidate set query, an active learning framework that reduces the labeling cost effectively and efficiently by narrowing down the candidate set likely to include the ground-truth class. We have also proposed a novel acquisition function that balances model performance and labeling cost by taking ex-

[2]For computational feasibility we therefore cap its AL budget at 5000 images and its calibration set size at 500.
[3]We adopt the accelerated code from Zhang et al. (2024).

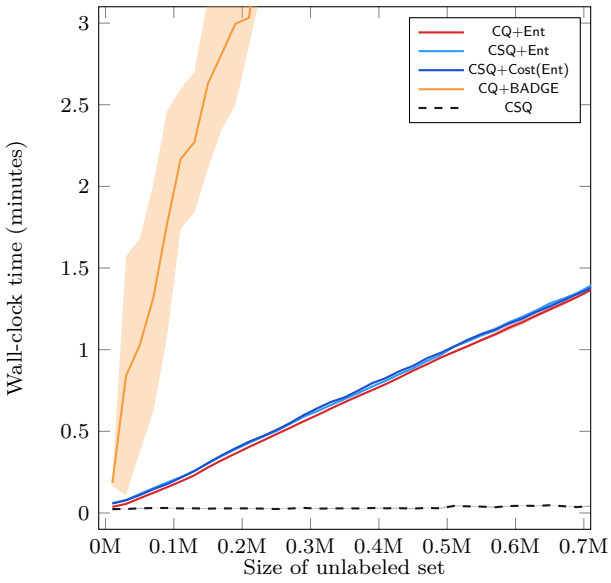


Figure 7: Wall-clock time versus unlabeled-pool size on ImageNet64x64. CSQ (dashed) runs in constant time, independent of pool size. Our sampling (Cost(Ent)) adds a negligible overhead, keeping the same linear complexity as plain entropy sampling.

Table 3: Detailed comparison of wall-clock time measured in *minutes* (mean ± standard deviation across three different random seeds) with an unlabeled data pool of size 1.21 million from ImageNet64x64. The table decomposes the total runtime into *Query* (for constructing candidate sets) and *Sampling* (for selecting informative samples). While conventional querying strategies (CQ+Ent and CQ+BADGE) incur no query-time cost by presenting all classes to the oracle, the proposed CSQ framework introduces a marginal overhead (<0.05 minutes), thereby maintaining practical efficiency. Also, our cost-efficient sampling Cost(Ent) maintains the linear time complexity of standard entropy sampling, while adding only marginal additional cost. Together, these results show that the full pipeline CSQ+Cost(Ent) achieves competitive computational efficiency without sacrificing scalability.

	Query	Sampling	Total
CQ+Ent	$0.00_{\pm 0.00}$	$2.29_{\pm 0.01}$	$2.29_{\pm 0.01}$
CQ+BADGE	$0.00_{\pm 0.00}$	$17.64_{\pm 0.31}$	$17.64_{\pm 0.31}$
CSQ+Ent	$0.04_{\pm 0.01}$	$2.27_{\pm 0.01}$	$2.31_{\pm 0.01}$
CSQ+Cost(Ent)	$0.04_{\pm 0.00}$	$2.26_{\pm 0.02}$	$2.31_{\pm 0.02}$

Figure 8: Examples of input images and their corresponding candidate sets constructed using our method at the fifth AL round on ImageNet64x64. For each image, the ✓ indicates the ground-truth class. Our method adjusts the size of each candidate set on the fly: it shrinks the set for *confident* cases (*left*) to reduce annotation cost, and enlarges it for *uncertain* ones (*right*) to include the ground-truth class.

pected candidate set sizes into account. Empirical evaluations on CIFAR-10, CIFAR-100, and ImageNet64x64 confirm the effectiveness of our framework.

Limitations and future work. One limitation is that the proposed acquisition function lacks theoretical guarantee for label complexity (Dasgupta, 2011; Hanneke et al., 2014) at this point. Establishing a theoretical understanding to quantify the cost required to achieve a target performance remains an interesting direction for future work.

References

Anastasios N Angelopoulos, Stephen Bates, et al. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *Proc. International Conference on Machine Learning (ICML)*, 2020.

Shilpa Arora, Eric Nyberg, and Carolyn Rose. Estimating annotation cost for active learning in a multi-annotator environment. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pp. 18–26, 2009.

Nabiha Asghar, Pascal Poupart, Xin Jiang, and Hang Li. Deep active learning for dialogue generation. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM)*, 2017.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *Proc. International Conference on Learning Representations (ICLR)*, 2020.

Yalong Bai, Yuxiang Chen, Wei Yu, Linfang Wang, and Wei Zhang. Products-10k: A large-scale product recognition dataset. *arXiv preprint arXiv:2008.10545*, 2020.

Aditya R Bhattacharya and Shayok Chakraborty. Active learning with n-ary queries for image recognition. In *Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2019.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021.

Seong Jin Cho, Gwangsu Kim, Junghyun Lee, Jinwoo Shin, and Chang D. Yoo. Querying easily flip-flopped samples for deep active learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=THUBTfSAS2`.

Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

Corinna Cortes. Support-vector networks. *Machine Learning*, 1995.

Timothee Cour, Ben Sapp, and Ben Taskar. Learning from partial labels. In *Journal of Machine Learning Research (JMLR)*, 2011.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 9268–9277, 2019.

Sanjoy Dasgupta. Two faces of active learning. *Theoretical computer science*, 412(19):1767–1781, 2011.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009a. doi: 10.1109/CVPR.2009.5206848.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: a large-scale hierarchical image database. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009b.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Pan Du, Suyun Zhao, Hui Chen, Shuwen Chai, Hong Chen, and Cuiping Li. Contrastive coding for active learning under class distribution mismatch. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8927–8936, 2021.

Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. Data determines distributional robustness in contrastive language image pre-training (clip). In *Proc. International Conference on Machine Learning (ICML)*, pp. 6216–6234. PMLR, 2022.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

Dominik Fuchsgruber, Tom Wollschläger, Bertrand Charpentier, Antonio Oroz, and Stephan Günnemann. Uncertainty for active learning on graphs. In *Forty-first International Conference on Machine Learning (ICML)*, 2024.

Guy Hacohen and Daphna Weinshall. How to select which active learning strategy is best suited for your specific problem and budget. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 13395–13407. Curran Associates, Inc., 2023.

Guy Hacohen, Avihu Dekel, and Daphna Weinshall. Active learning on a budget: Opposite strategies suit high and low budgets. In *International Conference on Machine Learning (ICML)*, pp. 8175–8195. PMLR, 2022.

Steve Hanneke et al. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Tao He, Xiaoming Jin, Guiguang Ding, Lan Yi, and Chenggang Yan. Towards better uncertainty sampling: Active learning with multiple views for deep convolutional neural network. In *IEEE International Conference on Multimedia and Expo (ICME)*, 2019.

Marek Herde, Denis Huseljic, Bernhard Sick, and Adrian Calma. A survey on cost types, interaction schemes, and annotator performance models in selection algorithms for active learning in classification. *IEEE Access*, 9:166970–166989, 2021.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Hengtong Hu, Lingxi Xie, Zewei Du, Richang Hong, and Qi Tian. One-bit supervision for image classification. *Proc. Neural Information Processing Systems (NeurIPS)*, 33:501–511, 2020.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Sheng-Jun Huang, Songcan Chen, and Zhi-Hua Zhou. Multi-label active learning: query type matters. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, volume 15, pp. 946–952, 2015.

Sehyun Hwang, Sohyun Lee, Sungyeon Kim, Jungseul Ok, and Suha Kwak. Combating label distribution shift for active domain adaptation. In *Proc. European Conference on Computer Vision (ECCV)*, pp. 549–566. Springer, 2022.

Sehyun Hwang, Sohyun Lee, Hoyoung Kim, Minhyeon Oh, Jungseul Ok, and Suha Kwak. Active learning for semantic segmentation with multi-class label query. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2023.

Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Breaking the interactive bottleneck in multi-class classification with active selection and binary feedback. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

Xin Kang, Xuefeng Shi, Yunong Wu, and Fuji Ren. Active learning with complementary sampling for instructing class-biased multi-label text emotion classification. *IEEE Transactions on Affective Computing*, 14(1):523–536, 2020.

Hoyoung Kim, Sehyun Hwang, Suha Kwak, and Jungseul Ok. Active label correction for semantic segmentation with foundation models. In *Proc. International Conference on Machine Learning (ICML)*, 2024.

Yoon-Yeong Kim, Youngjae Cho, JoonHo Jang, Byeonghu Na, Yeongmin Kim, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. Saal: sharpness-aware active learning. In *Proc. International Conference on Machine Learning (ICML)*, 2023.

Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision (CVPR)*, pp. 101–110, 2019.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Proc. Neural Information Processing Systems (NeurIPS)*, 32, 2019.

Suraj Kothawade, Nathan Beck, Krishnateja Killamsetty, and Rishabh Iyer. Similar: Submodular information measures based active learning in realistic scenarios. *Proc. Neural Information Processing Systems (NeurIPS)*, 34:18685–18697, 2021.

Akshay Krishnamurthy, Alekh Agarwal, Tzu-Kuo Huang, Hal Daumé III, and John Langford. Active learning for cost-sensitive classification. *Journal of Machine Learning Research (JMLR)*, 20(65):1–50, 2019.

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

David D. Lewis. Reuters-21578 text categorization test collection, 1997. URL `http://www.daviddlewis.com/resources/testcollections/reuters21578/`.

Yezheng Liu, Zhe Li, Chong Zhou, Yuanchun Jiang, Jianshan Sun, Meng Wang, and Xiangnan He. Generative adversarial active learning for unsupervised outlier detection. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1517–1528, 2019a.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proc. International Conference on Learning Representations (ICLR)*, 2019.

Sergio Matiz and Kenneth E Barner. Conformal prediction based active learning by linear regression optimization. *Neurocomputing*, 388:157–169, 2020.

Kun-Peng Ning, Xun Zhao, Yu Li, and Sheng-Jun Huang. Active learning for open-set annotation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41–49, 2022.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Natalia Ostapuk, Jie Yang, and Philippe Cudré-Mauroux. Activelink: deep active learning for link prediction in knowledge graphs. In *The World Wide Web Conference (WWW)*, 2019.

Dongmin Park, Yooju Shin, Jihwan Bang, Youngjun Lee, Hwanjun Song, and Jae-Gil Lee. Meta-query-net: Resolving purity-informativeness dilemma in open-set active learning. *Proc. Neural Information Processing Systems (NeurIPS)*, 35:31416–31429, 2022.

Buyue Qian, Xiang Wang, Fei Wang, Hongfei Li, Jieping Ye, and Ian Davidson. Active learning from relative queries. In *Proc. International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Yucheng Tang, Alan L Yuille, Zongwei Zhou, et al. Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021.

Yaniv Romano, Matteo Sesia, and Emmanuel Candes. Classification with valid and adaptive coverage. *Proc. Neural Information Processing Systems (NeurIPS)*, 33:3581–3591, 2020.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *Proc. International Conference on Learning Representations (ICLR)*, 2018.

Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

Burr Settles, Mark Craven, and Lewis Friedland. Active learning with real annotation costs. In *Proceedings of the NIPS workshop on cost-sensitive learning*, volume 1, 2008.

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

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proc. International Conference on Machine Learning (ICML)*, pp. 6105–6114. PMLR, 2019.

Toan Tran, Thanh-Toan Do, Ian Reid, and Gustavo Carneiro. Bayesian generative active deep learning. In *International conference on machine learning*, pp. 6295–6304. PMLR, 2019.

Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *Proc. International Conference on Machine Learning (ICML)*, ICML '99, pp. 444–453, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1558606122.

Byron C Wallace, Kevin Small, Carla E Brodley, and Thomas A Trikalinos. Who should label what? instance allocation in multiple expert active learning. In *Proceedings of the 2011 SIAM international conference on data mining*, pp. 176–187. SIAM, 2011.

Di Wang, Jing Zhang, Bo Du, Minqiang Xu, Lin Liu, Dacheng Tao, and Liangpei Zhang. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. *Proc. Neural Information Processing Systems (NeurIPS)*, 36, 2024.

Jing Wang, Jie Shen, Xiaofei Ma, and Andrew Arnold. Uncertainty-based active learning for reading comprehension. *Transactions on Machine Learning Research (TMLR)*, 2022. ISSN 2835-8856.

Zengmao Wang, Bo Du, Weiping Tu, Lefei Zhang, and Dacheng Tao. Incorporating distribution matching into uncertainty for multiple kernel active learning. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2019.

Zheng Wang and Jieping Ye. Querying discriminative and representative samples for batch mode active learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2015.

Ming-Kun Xie and Sheng-Jun Huang. Multi-label learning with pairwise relevance ordering. *Proc. Neural Information Processing Systems (NeurIPS)*, 34:23545–23556, 2021.

Yang Yang, Yuxuan Zhang, Xin Song, and Yi Xu. Not all out-of-distribution data are harmful to open-set active learning. *Proc. Neural Information Processing Systems (NeurIPS)*, 36, 2024.

Ofer Yehuda, Avihu Dekel, Guy Hacohen, and Daphna Weinshall. Active learning through a covering lens. *Proc. Neural Information Processing Systems (NeurIPS)*, 35:22354–22367, 2022.

Guoxian Yu, Xia Chen, Carlotta Domeniconi, Jun Wang, Zhao Li, Zili Zhang, and Xiangliang Zhang. Cmal: Cost-effective multi-label active learning by querying subexamples. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 34(5):2091–2105, 2020.

Sergey Zagoruyko. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12104–12113, 2022.

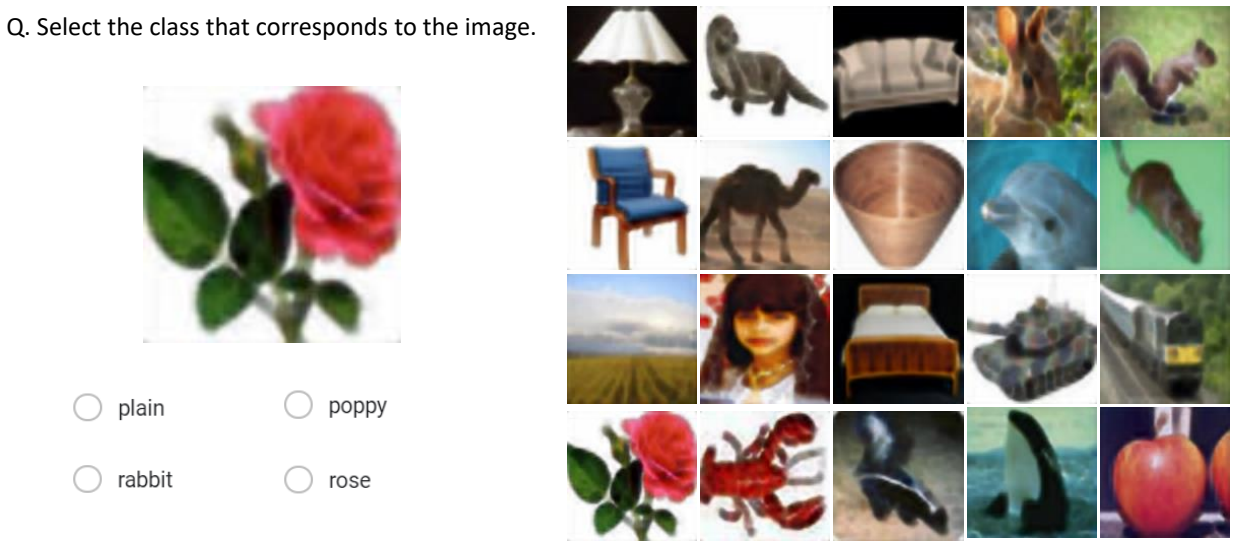
Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2018.

Jifan Zhang, Yifang Chen, Gregory Canal, Arnav Mohanty Das, Gantavya Bhatt, Stephen Mussmann, Yinglun Zhu, Jeff Bilmes, Simon Shaolei Du, Kevin Jamieson, and Robert D Nowak. Labelbench: A comprehensive framework for benchmarking adaptive label-efficient learning. *Journal of Data-centric Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=Y2QcZfwHE7`. Reproducibility Certification.

Yuhang Zhang, Xiaopeng Zhang, Lingxi Xie, Jie Li, Robert C Qiu, Hengtong Hu, and Qi Tian. One-bit active query with contrastive pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9697–9705, 2022.

Appendix

A Details of user study



(a) Questionnaire with four candidates ($k = 4$) (b) Example queries in CIFAR-100

Figure 9: Questionnaire and examples used in the user study. (a) Each question contains an instruction, an image, and a set of candidates. In this case, the candidate set size is 4. (b) We utilize 20 images from CIFAR-100, each with a resolution of 128 x 128 pixels.

We conduct a user study to examine how the size of a candidate set, k in Sec. 3.1, affects the annotation time in practice. Figure 9 presents examples of the questionnaire and all images used in our user study. To facilitate easy comparison with the theoretical costs (Hu et al., 2018), we set the candidate set sizes to 4, 8, 16, and 32. To be specific about Figure 9, we use CIFAR-100 images resized to 128×128 using super resolution[4] to enhance visibility for annotators. We first randomly select 20 classes in CIFAR-100 and choose one image per class to organize the questionnaires. For small-sized candidate sets, we ensure the inclusion of the ground truth by randomly trimming around it when generating the candidate sets.

We divide 44 annotators into four groups of 11 for each candidate set size to perform labeling tasks. To account for potential outliers, we exclude the results of the annotators whose time taken deviates the most from the average time in each group. Table 4 shows that as the candidate set size increases, the time per query increases and the accuracy decreases. In addition, on the right side of Table 4, the comparison between the experimental costs and theoretical costs reveals a significant correlation of 0.97.

Table 4: User study for different sizes of candidate set query.

k	Total time (s)	Time per query (s)	Accuracy (%)	Experimental	Theoretical
4	$\mathbf{69.4}_{\pm 13.8}$	$\mathbf{3.47}_{\pm 0.69}$	$\mathbf{100.0}_{\pm 0.0}$	2.0	2
8	$91.5_{\pm 27.3}$	$5.20_{\pm 1.36}$	$98.5_{\pm 3.2}$	2.6	3
16	$116.9_{\pm 29.6}$	$6.94_{\pm 1.48}$	$99.5_{\pm 1.5}$	3.4	4
32	$166.9_{\pm 30.8}$	$8.35_{\pm 1.54}$	$95.5_{\pm 5.2}$	4.8	5

[4]https://www.kaggle.com/datasets/joaopauloschuler/cifar100-128x128-resized-via-cai-super-resolution

B Implementation details and configuration

Table 5 presents the configuration of our main experiments for each dataset. In all experiments, we fixed the per-round budget, which limits the number of annotated instances per active learning (AL) round. Given this budget constraint, we compute the labeling cost for each AL round to assess labeling efficiency. The batch size for CIFAR-10, CIFAR-100, and ImageNet64x64 was determined to be 128. We normalized the input image to ensure the stability of the training. We trained our classification model on CIFAR-10 and CIFAR-100 using NVIDIA RTX 3090 and on ImageNet64x64 using 4 NVIDIA A100 GPUs in parallel. The training requires about 5 GPU hours for CIFAR-10 and CIFAR-100, and about 1.5 GPU days for ImageNet64x64.

Table 5: Detailed dataset and budget configuration for the proposed scenario.

Dataset	L	$\log_2 L$	Size	Cost of full label	# of rounds	Per-round budget
CIFAR-10	10	3.322	50K	166.1K	10	6K
CIFAR-100	100	6.644	50K	332.2K	9	6K
ImageNet64x64	1000	9.966	1.2M	12.7M	16	60K

Code. This part demonstrates the reproducibility of our work by providing comprehensive details on the source code release. We have made available the entire framework, which includes the data sampling methods, evaluation procedures, and the overall training pipeline. Our aim is to ensure that other researchers can easily replicate and build upon our results. To get started with running the code, please refer to the `script.sh` and `README.md` files. `README.md` contains the instructions to comprehend and execute our experiments seamlessly, and `script.sh` includes some example commands. To understand our proposed method better, you can examine the Python script `al/strategy_dtopk.py`. This file includes the implementation details of our active learning strategies, particularly *candidate set query* design. Furthermore, our code can run on CIFAR-10, CIFAR-100 [5], and ImageNet64x64 [6], which are available online. Note that you can modify the running configuration such as dataset, sampling method, and budget through command-line arguments.

C Additional clarification on candidate set construction

The detailed procedure of computing $\hat{Q}(\alpha)$ in Eq. (4). We begin with computing the collection of conformal scores $\mathbf{s}$ for the calibration dataset $\mathcal{D}_{\text{cal}}$. For each data point $(\mathbf{x}_i, y_i) \in \mathcal{D}_{\text{cal}}$, the conformal score is defined as:

$$s_i := 1 - P_{\boldsymbol{\theta}}(y_i \mid \mathbf{x}_i), \quad \text{for } i = 1, 2, \cdots, n_{\text{cal}} \,, \tag{9}$$

where $n_{\text{cal}} = |\mathcal{D}_{\text{cal}}|$. Using these scores, we define the empirical distribution function $F_n(s)$, which measures the proportion of scores less than or equal to a given value s. Formally, $F_n(s)$ is expressed as:

$$F_n(s) = \frac{1}{n_{\text{cal}}} \sum_{i=1}^{n_{\text{cal}}} \mathbb{1}[s_i \leq s] \,, \tag{10}$$

where $\mathbb{1}[\cdot]$ is an indicator function. The $(1-\alpha)$ empirical quantile is then defined as the smallest score s_i such that the proportion of scores satisfying $s_i \leq s$ is at least $(1-\alpha)$. Mathematically, this is given as $\min_{i \in [n_{\text{cal}}]} \{F_n(s_i) \geq 1-\alpha\}$, where $[n_{\text{cal}}] = \{1, 2, \cdots, n_{\text{cal}}\}$.

$$\hat{Q}(\alpha) := \min_{i \in [n_{\text{cal}}]} \{F_n(s_i) \geq (1-\alpha)\} \,. \tag{11}$$

Note that Eq. (11) is equivalent to Eq. (4).

[5] `https://www.cs.toronto.edu/~kriz/cifar.html`
[6] `https://patrykchrabaszcz.github.io/Imagenet32/`

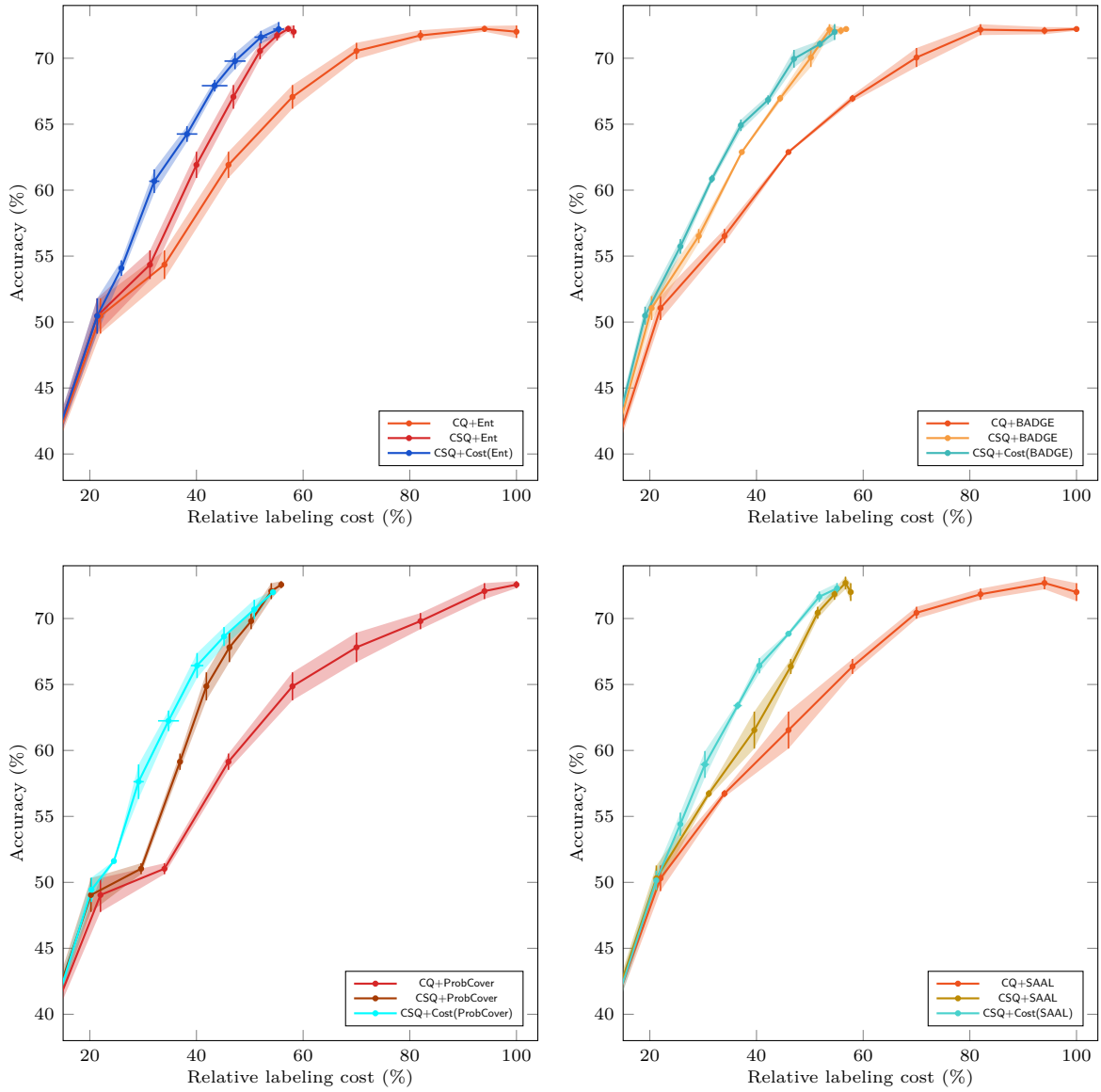
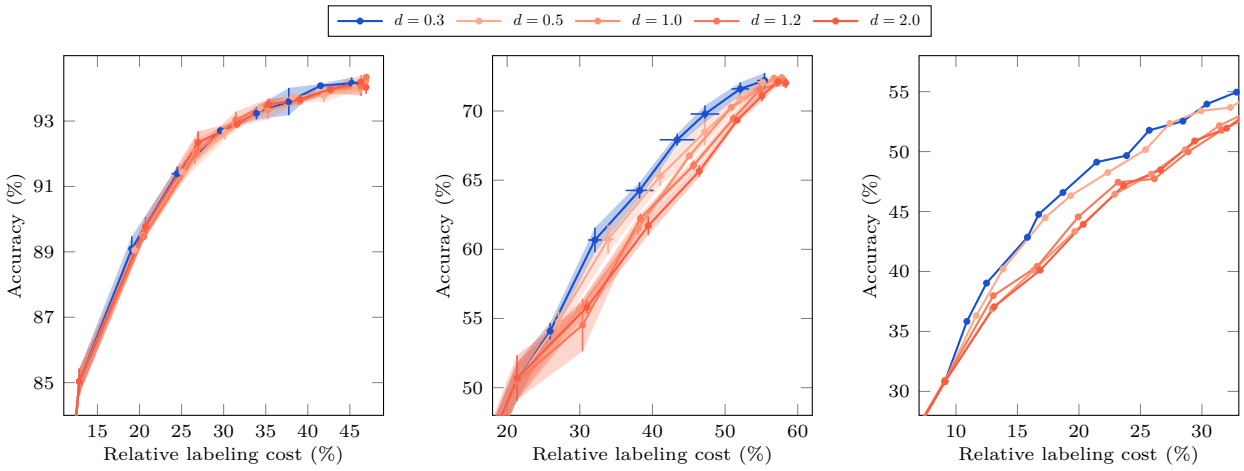


Figure 10: Comparison of different sampling methods and their cost-sampling variants on CIFAR-100. Each subplot shows a triple of corresponding methods.

D Impact of proposed cost-efficient sampling across different sampling strategies

In Figure 10, we compare different sampling strategies when combined with their cost-efficient variants on CIFAR-100. Notably, the performance of cost-efficient sampling consistently improves when combined with a range of acquisition functions. These results show that our cost-efficient acquisition method (Eq. (8)) performs consistently well when paired with four diverse acquisition strategies: entropy sampling, BADGE (Ash et al., 2020), ProbCover (Yehuda et al., 2022), and SAAL (Kim et al., 2023). This suggests that it can be readily combined with a broad range of acquisition functions beyond those evaluated in our experiments.



(a) CIFAR-10 (b) CIFAR-100 (c) ImageNet64x64

Figure 11: Accuracy (%) versus relative labeling cost (%) with varying hyperparameter d in Eq. (8) across AL rounds, evaluated on CIFAR-10, CIFAR-100 and ImageNet64x64 with CSQ+Cost(Ent). We set $d = 0.3$ for all datasets in our main experiments.

E Impact of informativeness-cost balancing hyperparameter d

The hyperparameter d in our acquisition function (Eq. (8)) balances the trade-off between labeling cost and the informativeness of a sample, requiring both factors to be considered. We provide a comprehensive analysis showing the trend of performance in accuracy with varying d values over AL rounds on CIFAR-10, CIFAR-100, and ImageNet64x64 in Figure 11. For CIFAR-10 (Figure 11a), both accuracy and labeling cost remain robust to the change of d, varying only 0.5%p in accuracy. For CIFAR-100 and ImageNet64x64 (Figure 11b and Figure 11c), the overall performance improves as d decreases. Overall, the performance remains robust for lower values of d, showing little sensitivity across different datasets. This suggests that d can be easily selected without extensive tuning. We fix $d = 0.3$ in all main experiments to demonstrate the robustness of our method without dataset-specific adjustment.

F Discussion on handling outliers and anomalous datapoints

Dealing with out-of-distribution (OOD) data points showing high uncertainty scores has been a chronic issue in active learning and may affect the efficiency of candidate set query (CSQ). Recent open-set active learning approaches (Du et al., 2021; Kothawade et al., 2021; Ning et al., 2022; Park et al., 2022; Yang et al., 2024) tackle this by filtering out OOD samples during active sampling using an OOD classifier. Our CSQ framework integrates seamlessly with these methods, focusing on labeling in-distribution (ID) samples to prevent cost inefficiency.

However, as OOD classifiers are not flawless, some OOD samples may still be selected. One advantage of our method is its ability to leverage the calibration set to capture information about such mixed OOD samples. This enables adjustments such as increasing the OOD classifier threshold to exclude more OOD-like data or incorporating the OOD ratio into the alpha optimization process in Eq. (7). Optimizing the combination of OOD and ID classifier scores within the calibration set or designing better OOD-aware queries presents promising future research directions.

G Compatibility between candidate set construction and uncertain samples

Figure 12 compares CSQ and conventional query (CQ) on CIFAR-100 with entropy-based sampling (Ent) and our acquisition function with entropy measure (Cost(Ent), Eq. (8)) across AL rounds, with a fixed number of samples per round.

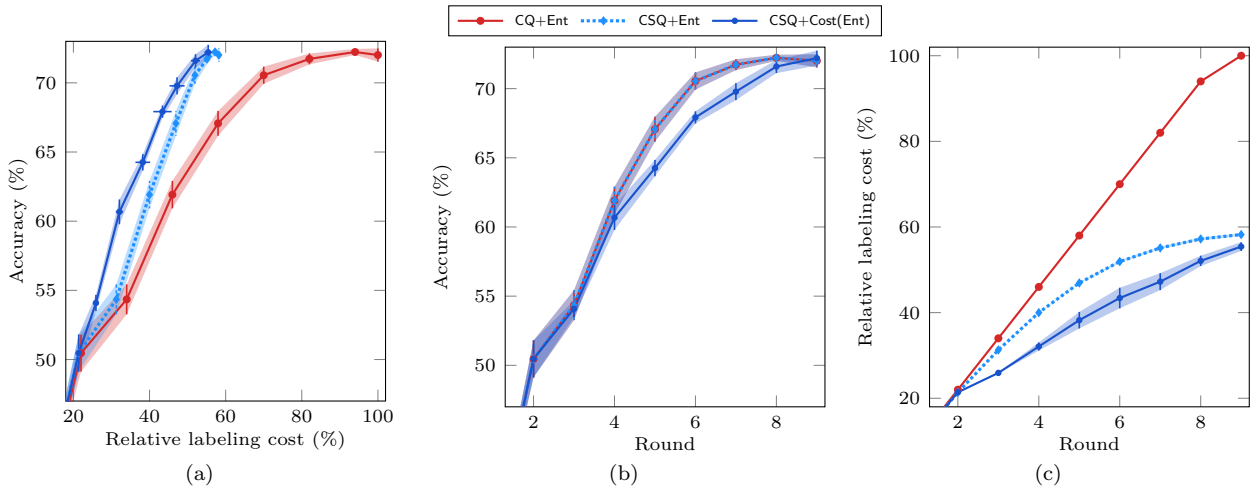


Figure 12: Comparison of candidate set query (CSQ) and conventional query (CQ) on CIFAR-100 with entropy sampling (Ent) and cost-efficient entropy sampling (Cost(Ent)) varying AL rounds. A fixed number of samples are selected at each AL round. (a) Accuracy (%) versus relative labeling cost (%) showing the accuracy per cost. (b) Accuracy (%) versus AL rounds showing the accuracy varies with the number of samples. Note that the lines of CQ+Ent and CSQ+Ent completely overlap, as they use the same sampling method. (c) Relative labeling cost (%) versus AL rounds.

Our acquisition function provides superior accuracy per cost. The comparison between CSQ+Cost(Ent) and CSQ+Ent demonstrates that the proposed acquisition function reduces labeling costs with only a marginal accuracy trade-off.

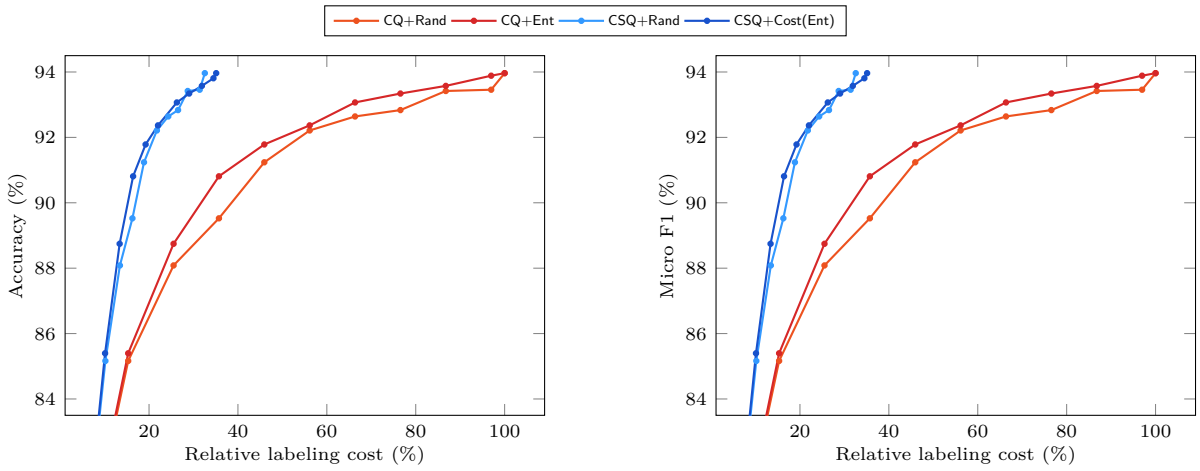
Candidate set query (CSQ) can reduce labeling costs even for uncertain samples. The comparison between CQ+Ent and CSQ+Ent demonstrates that CSQ effectively reduces labeling costs, even with uncertainty-based sampling methods like entropy sampling. This shows that CSQ can narrow down annotation options even for uncertain samples. Note that CSQ+Ent shows the same accuracy as CQ+Ent, since they use the same sampling method.

H Experiments in language domain

Dataset. The R52 dataset (Lewis, 1997) is a subset of the Reuters-21578 (Lewis, 1997) news collection, specifically curated for text classification tasks. It comprises documents categorized into 52 distinct classes, with a total of 9,130 documents. The dataset is divided into 6,560 training documents and 2,570 testing documents. Each document is labeled with a single category, and the categories are selected to ensure that each has at least one document in both the training and testing sets. This structure makes the R52 dataset particularly suitable for evaluating text classification models.

Implementation details. We adopt an SVM classifier (Cortes, 1995) with sigmoid kernel for classification. We conduct 11 AL rounds of consecutive data sampling and model updates, where the per-round budget is 600. The hyperparameter d for our acquisition function is set as 1.2. In the initial round, we randomly sample 300 samples. In each round, the model is evaluated based on three factors: its accuracy (%) and Micro-F1 (%).

Figure 13 presents a comparison of candidate set query (CSQ) and conventional query (CQ) on the text classification dataset (R52) with random sampling (Rand), entropy sampling (Ent), and our acquisition function with entropy measure (Cost(Ent), Eq. (8)) across AL rounds. CSQ approaches consistently outperform the CQ baselines by a significant margin across various budgets and acquisition functions. Especially at round 10, CSQ+Rand reduces labeling cost by 65.6%p compared to its conventional query baseline. The result demonstrates that the proposed CSQ framework generalizes to the text classification domain.



(a) Accuracy (b) Micro-F1

Figure 13: Comparison between conventional query (CQ) and candidate set query (CSQ) with random sampling (Rand), entropy sampling (Ent), and cost-efficient entropy sampling (Cost(Ent)) on text classification task with R52 dataset. (a) Accuracy (%) versus relative labeling cost (%). (b) Micro-F1 (%) versus relative labeling cost (%). CSQ approaches (blue lines) consistently outperform the CQ baselines (red lines) by a significant margin across various budgets and acquisition functions.

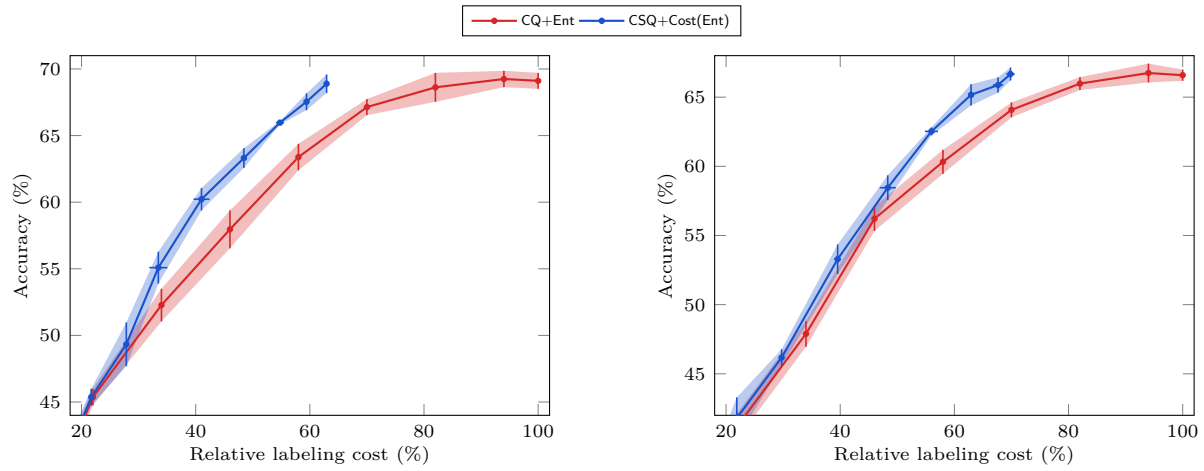
I Experiments on real-world datasets

Experiment on datasets containing label noise. We evaluate the candidate set query (CSQ) framework on CIFAR-100 with noisy labels, simulating a scenario where human annotators misclassify images into random classes with a noise rate ϵ. This is modeled using a uniform label noise (Frénay & Verleysen, 2013) with ϵ set to 0.05 and 0.1. Note that this scenario is unfavorable for CSQ, as a misclassifying annotator would reject the actual true label even if the candidate set includes it. Figure 14 compares CSQ and conventional query (CQ) on CIFAR-100 with noisy labels using entropy sampling (Ent) and our acquisition function with entropy measure (Cost(Ent)) across 2, 6, and 9 rounds.

Despite the disadvantageous scenario, our method (CSQ+Cost(Ent)) reduces labeling cost compared to the baseline (CQ+Ent) across varying AL rounds and noise rates. At round 9, CSQ+Cost(Ent) achieves cost reductions of 33.4%p and 27.4%p at noise rates of 0.05 and 0.1, respectively. It also consistently outperforms the baseline in terms of accuracy per labeling cost, demonstrating the robustness of CSQ. Additionally, CSQ has the potential to reduce label noise, as narrowing the candidate set can lead to more precise annotations. Our user study (Table 1) shows that reducing candidate set size improves annotation accuracy, suggesting that CSQ can further enhance performance by reducing label noises.

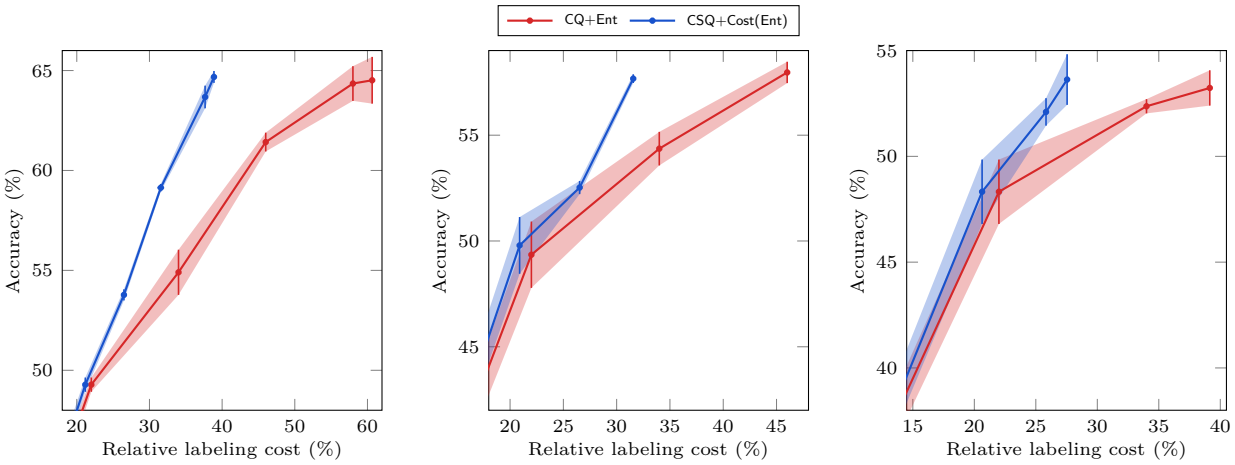
Experiment on datasets containing class imbalances. Figure 15 compares candidate set query (CSQ) and conventional query (CQ) on CIFAR-100-LT (Cui et al., 2019), a class-imbalanced version of CIFAR-100, using entropy sampling (Ent), and our acquisition function with entropy measure (Cost(Ent)) across AL rounds. The experiments use imbalance ratios (*i.e.*, ratios between the largest and smallest class sizes) of 3, 6, and 10. Note that the maximum AL rounds vary with the imbalance ratio due to dataset size, with a maximum of 4 rounds for ratios of 3 and 6, and 6 rounds for a ratio of 10.

The result shows that our method (CSQ+Cost(Ent)) reduces labeling cost compared to the baselines (CQ+Ent) by significant margins across varying AL rounds and imbalance ratios. Specifically, at round 4, CSQ+Cost(Ent) achieves cost reductions of 31.1%p and 29.2%p at imbalance ratios of 6 and 10, respectively. In terms of accuracy per labeling cost, CSQ+Cost(Ent) consistently outperforms the baseline, demonstrating the robustness of the CSQ framework in class-imbalanced scenarios.



(a) $\epsilon = 0.05$ (b) $\epsilon = 0.1$

Figure 14: Comparison between conventional query (CQ) and candidate set query (CSQ) with entropy sampling (Ent) and the proposed acquisition function with entropy measure (Cost(Ent)) on CIFAR-100 with label noise across AL rounds with varying noise level: (a) Noise rate of 0.05. (b) Noise rate of 0.1. The proposed CSQ+Cost(Ent) consistently outperforms CSQ+Ent across various AL rounds and noise rates.



(a) $\rho = 3$ (b) $\rho = 6$ (c) $\rho = 10$

Figure 15: Comparison between conventional query (CQ) and candidate set query (CSQ) with entropy sampling (Ent) and the proposed acquisition function with entropy measure (Cost(Ent)) on CIFAR-100-LT, a variant of CIFAR-100 with class imbalance, across AL rounds with varying imbalance level: (a) Imbalance ratio of 3. (b) Imbalance ratio of 6. (c) Imbalance ratio of 10. The proposed approach (CSQ+Cost(Ent)) consistently outperforms the baseline (CSQ+Ent) across various AL rounds and noise rates. Note that the maximum AL rounds vary with the imbalance ratio due to dataset size, with a maximum of 4 rounds for ratios of 3 and 6, and 6 rounds for a ratio of 10.

J Additional related work

Efficient query design. Designing efficient annotation queries reduces the annotation costs of crafting datasets. In AL, diverse types of queries have been investigated, including conventional classification queries (Huang et al., 2015; Kang et al., 2020; Yu et al., 2020; Xie & Huang, 2021; Cour et al., 2011), one-bit queries (Hu et al., 2020; Joshi et al., 2010) asking for yes or no answers, multi-class queries (Hwang et al., 2023) identifying all classes within a set of multiple instances, relative quires (Qian et al., 2013) asking for similarity of triplets, and correction queries (Kim et al., 2024) utilizing pseudo labels from the model. While

these query methods require tailored loss functions, our candidate set query (CSQ) is cost-efficient and provides complete supervision, integrating seamlessly with existing loss functions. The approach closely related to CSQ is the n-ary query (Bhattacharya & Chakraborty, 2019), which reduces the search space by asking for the correct class among top-n predictions of the model. However, the n-ary query uses a fixed number of top-n predictions for all data without considering individual sample difficulty. CSQ, on the other hand, adjusts the candidate set size based on sample difficulty and model performance using conformal prediction. Through rigorous comparisons, we demonstrate that CSQ achieves a superior model performance at the same cost compared to the previous query designs.

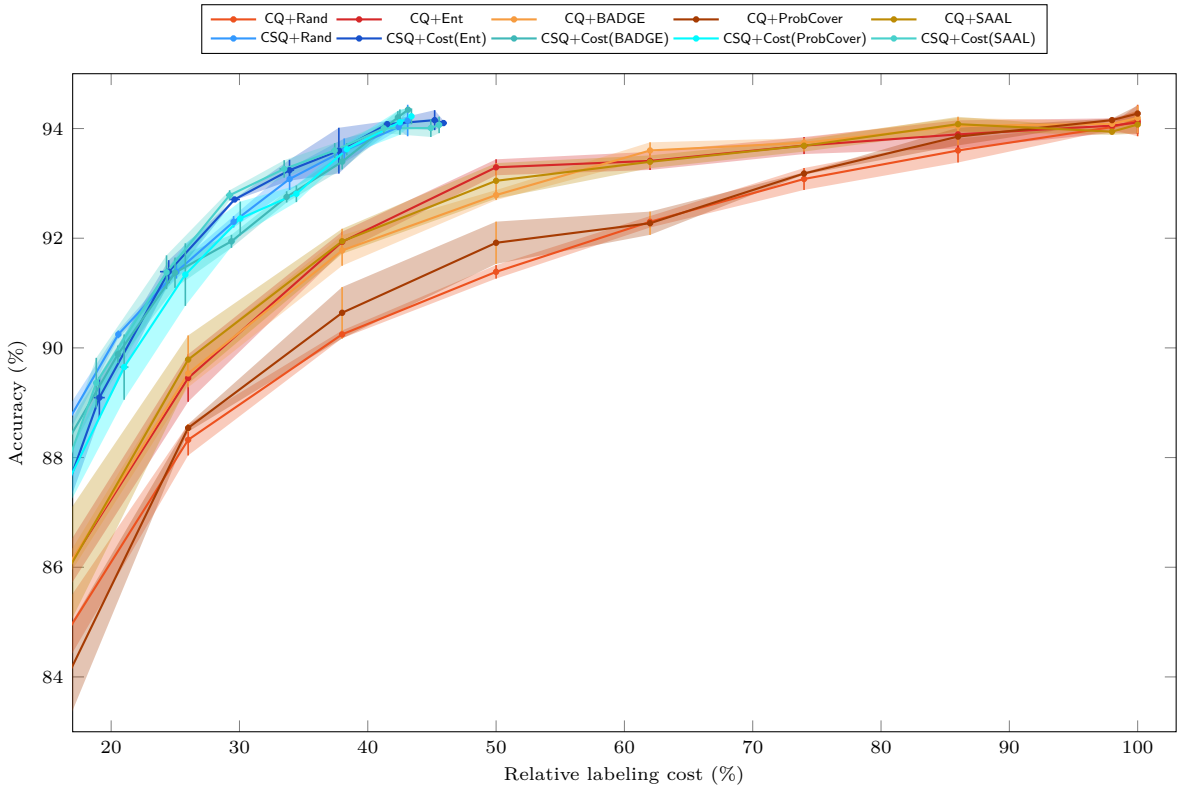


Figure 16: Accuracy (%) versus relative labeling cost (%) on CIFAR-10.

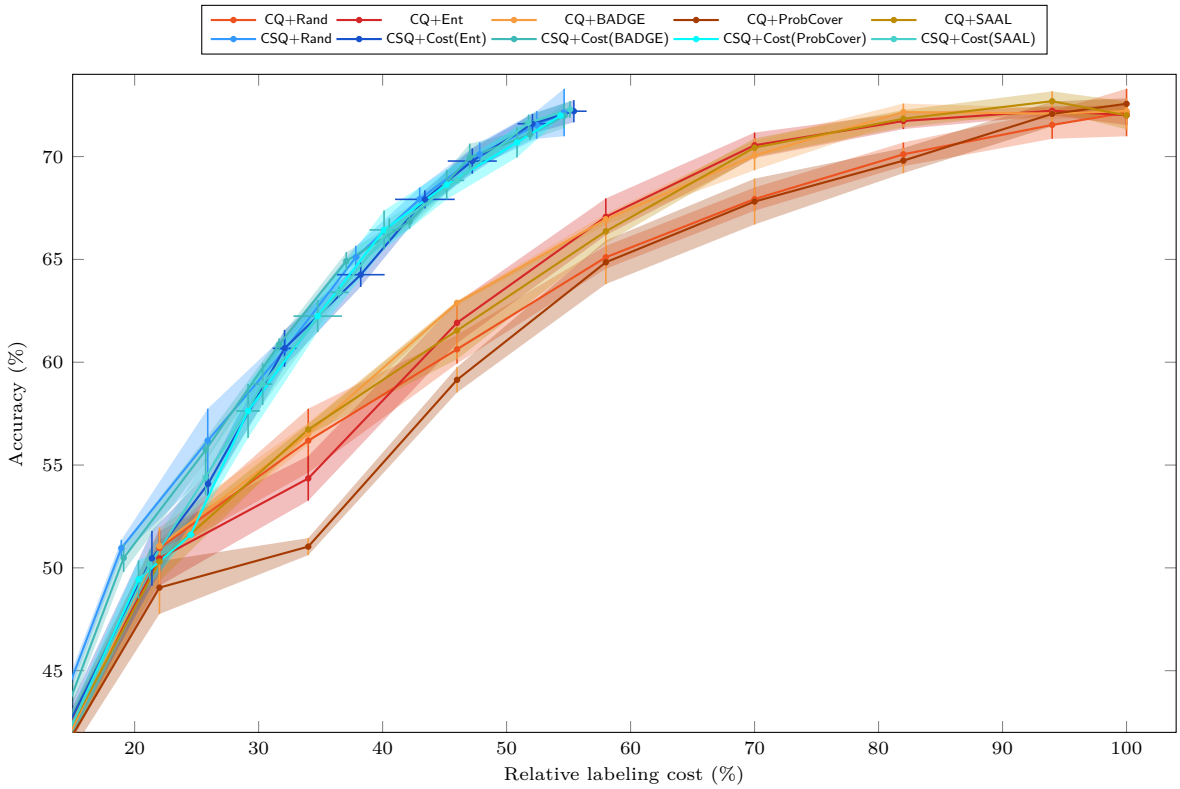
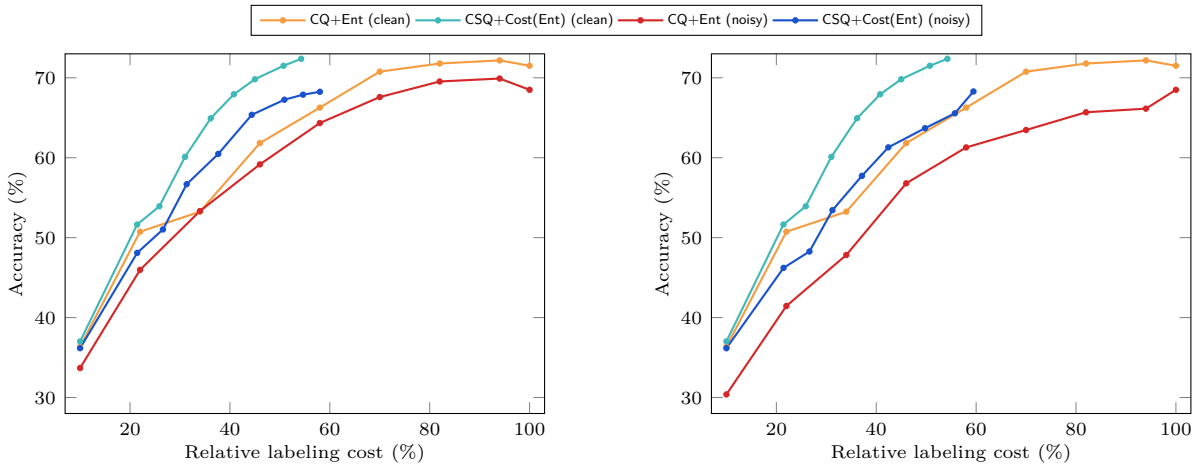
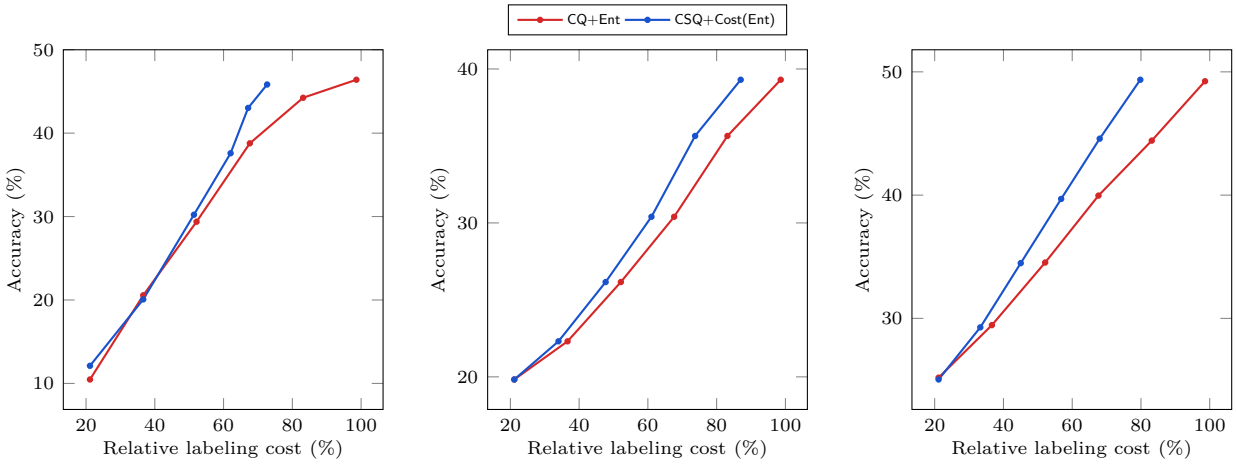


Figure 17: Accuracy (%) versus relative labeling cost (%) on CIFAR-100.



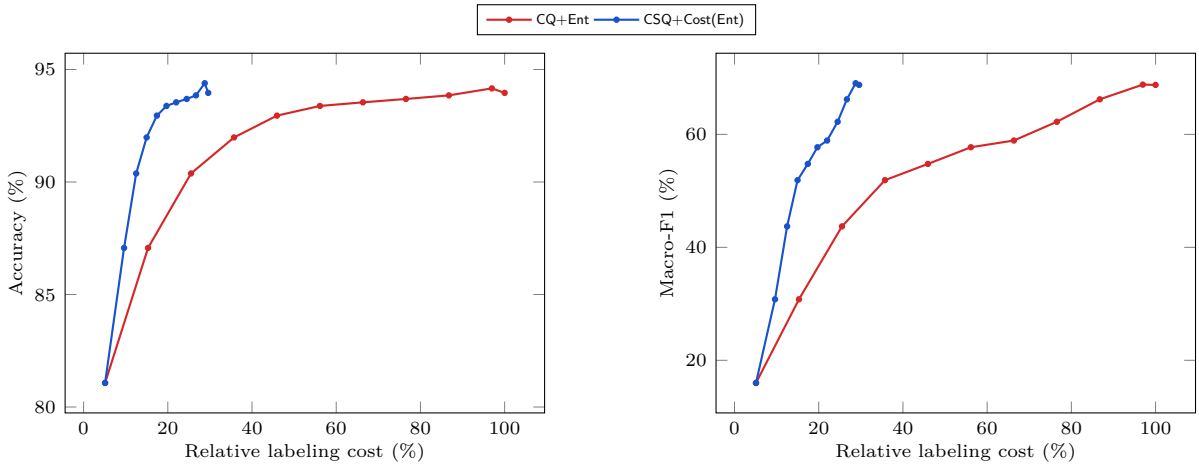
(a) $\epsilon = 0.05$ (b) $\epsilon = 0.1$

Figure 18: Comparison between conventional query combined with entropy sampling (CQ+Ent) and candidate set query combined with cost-efficient entropy sampling (CSQ+Cost(Ent)) on CIFAR-100 with label noise across AL rounds with varying noise level: (a) noise rate of 0.05, and (b) noise rate of 0.1. The label noise is generated by assigning a random label within candidate sets.



(a) EfficientNet-B3 (from scratch) (b) ResNet50 (DINO-pretrained) (c) ViT-S/16 (DINO-pretrained)

Figure 19: Comparison between conventional query combined with entropy sampling (CQ+Ent) and candidate set query combined with cost-efficient entropy sampling (CSQ+Cost(Ent)) on Products-10K (Bai et al., 2020) (a) EfficientNet-B3 (Tan & Le, 2019) is trained from scratch and evaluated on Products-10K. (b) ResNet50 (He et al., 2016) model pretrained with DINO (Caron et al., 2021) on ImageNet-1K (Deng et al., 2009b) is evaluated on Products-10K via linear probing. (c) ViT-S/16 (Dosovitskiy et al., 2021) model pretrained with DINO on ImageNet-1K is evaluated on Products-10K via linear probing.



(a) Accuracy (b) Macro-F1

Figure 20: Comparison between conventional query combined with entropy sampling (CQ+Ent) and candidate set query combined with cost-efficient entropy sampling (CSQ+Cost(Ent)) on the R52 dataset (Lewis, 1997). We adopt RoBERTa-Large (Liu et al., 2019b) as a frozen feature extractor. (a) Accuracy (%) and (b) Macro-F1 (%) versus relative labeling cost (%).

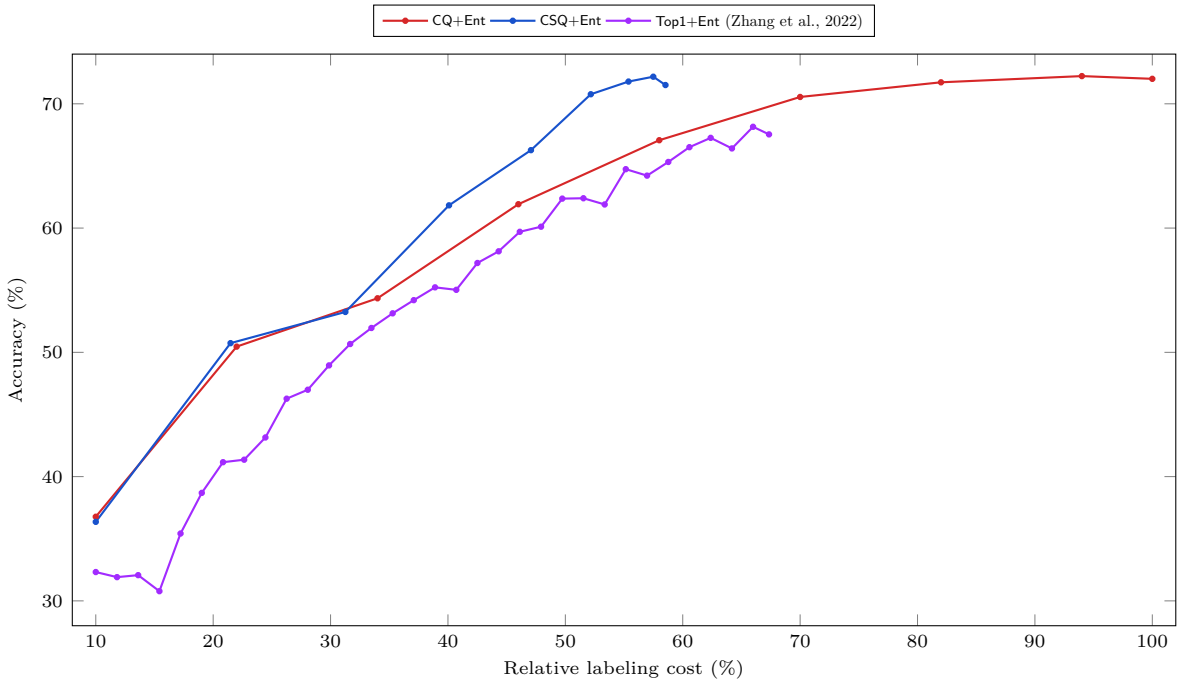
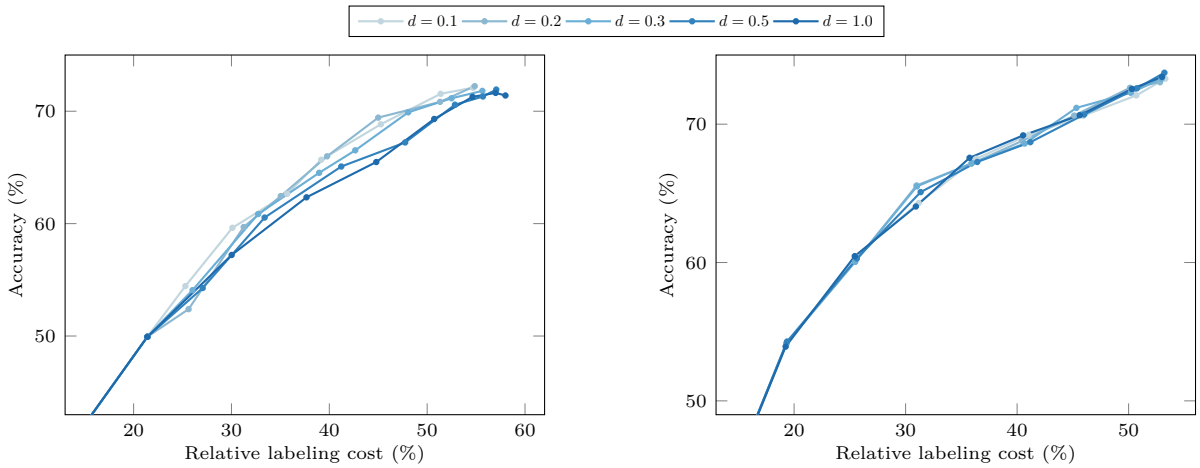
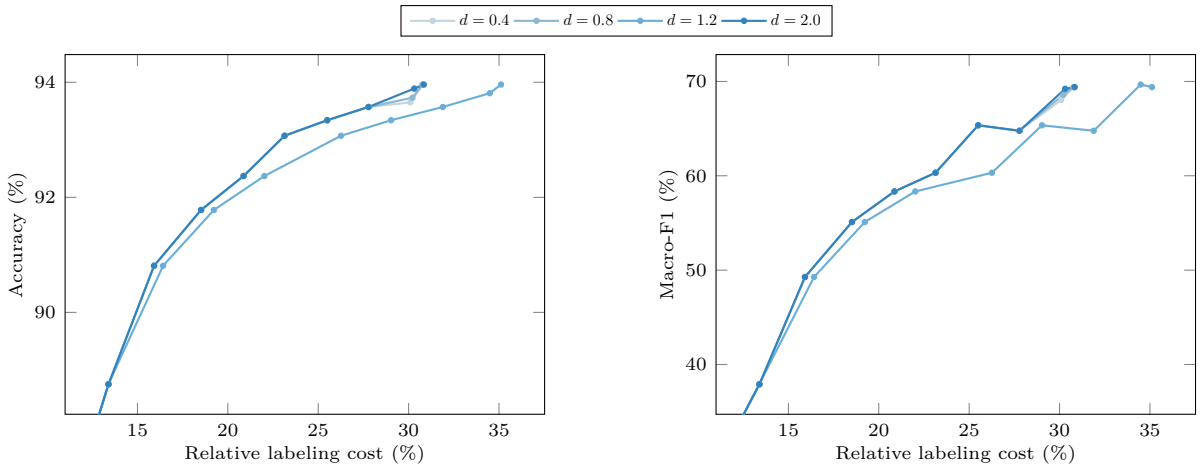


Figure 21: Comparison of conventional query (CQ), top-1 query (Top1) (Zhang et al., 2022), and candidate set query (CSQ) on CIFAR-100. Top1 asks the oracle whether the top-1 prediction is the ground-truth class, potentially yielding partial labels. Therefore, we employ negative learning loss (Kim et al., 2019) for the strategy and select samples with partial labels with replacement until they get full labels.



(a) CSQ+Cost(Ent) (b) CSQ+Cost(BADGE)

Figure 22: Accuracy (%) versus relative labeling cost (%) with varying hyperparameter d in Eq. (8) across AL rounds on CIFAR-100. (a) cost-efficient sampling combined with entropy sampling (CSQ+Cost(Ent)). (b) cost-efficient sampling combined with BADGE (CSQ+Cost(BADGE)).



(a) Accuracy (b) Macro-F1

Figure 23: (a) Accuracy (%) and (b) Macro-F1 (%) versus relative labeling cost (%) with varying hyperparameter d in Eq. (8) across AL rounds on R52 with CSQ+Cost(Ent).