Readiness-Aware Sample Selection for Noisy Labels with Class Imbalance

Chihyeon Choi[1], **Sangho Lee**[2], **Jiho Hong**[1], **Youngdoo Son**[1]*, **Hyungrok Do**[3]
[1]Department of Industrial and Systems Engineering, Dongguk University-Seoul
[2]School of Industrial and Systems Engineering, Gyeongsang National University
[3]Department of Population Health, NYU Grossman School of Medicine

Abstract

Noisy labels can cause deep neural networks to overfit incorrect annotations, leading to biased predictions and poor generalization. A common remedy is to select clean samples after a short warm-up phase, under the assumption that the model begins to learn class-specific patterns from the early training stage. However, this assumption often fails under class imbalance, as the majority classes can hinder the model from learning distinctive features of minority classes. In this paper, we propose a readiness-aware sample selection strategy that identifies clean samples only from classes whose distinctive features have been sufficiently learned and are thus considered ready for selection. We further introduce a novel negative learning scheme to enhance class separability by discouraging confusion with the most similar incorrect classes. The proposed method is supported by theoretical analysis and demonstrates outstanding performance on both benchmark and real-world datasets, showing improved robustness and generalization under noisy and class-imbalanced conditions.

1 Introduction

Deep neural networks (DNNs) have achieved remarkable success, largely driven by the availability of large-scale and well-annotated datasets [11, 9, 24, 28]. However, annotating high-quality labels is both time-consuming and expensive. This has led researchers to explore more cost-effective alternatives, such as web crawling or crowd-sourcing [15]. While these methods reduce annotation costs, they often introduce noisy labels, incorrect or unreliable annotations in some instances [13, 33, 42].

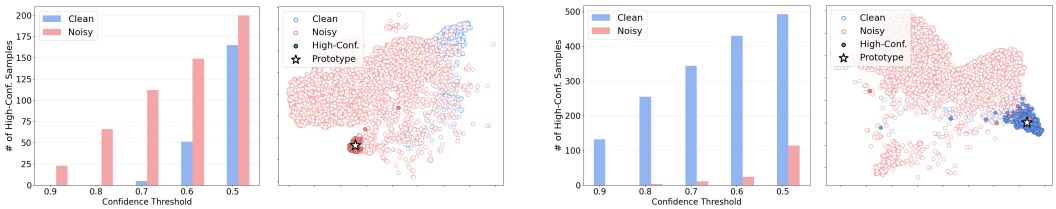
Noisy labels can cause DNNs to overfit incorrect labels, resulting in biased predictions and poor generalization. Consequently, a variety of robust learning algorithms have been proposed to ensure reliable training and effective performance even in the presence of noisy labels [7, 16, 35].

Among various approaches, sample selection methods have shown promising results in mitigating the impact of noisy labels. These methods aim to identify correctly annotated clean samples and treat them as labeled data, while treating the remaining samples as unlabeled [16, 27]. Most sample selection methods that have demonstrated strong performance rely on a small-loss criterion, which assumes that instances with lower training loss are more likely to be correctly labeled.

However, the effectiveness of the small-loss criterion depends on the implicit assumption that the dataset is class-balanced, that is, all classes have a similar number of samples. In the real world, this assumption often does not hold, as many datasets exhibit an imbalanced class distribution in which the number of instances varies across classes [37]. In such imbalanced datasets, samples from the majority classes tend to have smaller loss values [8]. Therefore, the small-loss criterion may lead to selecting the samples from the majority classes as the clean ones, whereas those of the minority

*Corresponding author (youngdoo@dongguk.edu)

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) After the warm-up phase

(b) After 10 epochs of training with RASS

Figure 1: In each figure, the left side shows the number of clean and noisy instances in high-confidence samples. We regard a sample as high-confidence when its confidence (predicted probability for the annotated class) exceeds the product of the maximum confidence among the samples in its class and a confidence threshold. The right side presents the feature representations of clean and noisy samples, along with the prototype, obtained by averaging the features of high-confidence samples when the confidence threshold is 0.8. The analysis focuses on class 6 from the CIFAR-10 dataset, which is simulated to have label noise and class imbalance (full results are provided in Appendix A). While the prototype aligns with noisy samples after the warm-up phase, the prototype becomes representative of clean samples after 10 epochs of training with the proposed RASS.

classes are more likely to be left as noisy. This suggests that the direct application of most existing sample selection methods may be inappropriate in scenarios with imbalanced data.

To address this limitation, recent studies have employed feature representations for sample selection under class imbalance, as they are relatively robust to imbalanced datasets [19, 39]. A common approach in these studies is to regard the samples close to the prototype in the feature space as clean [31, 41, 48]. However, existing methods often overlook that the model struggles to obtain representative prototypes for minority classes, particularly when they are dominated by mislabeled samples from majority classes.

Figure 1(a) illustrates a potential failure case of prototype-based approaches. In this example, a sample is regarded as high-confidence if its confidence score exceeds the product of the maximum confidence within its class and a predefined confidence threshold (e.g., 0.8). The class prototype is then computed by averaging the features of these high-confidence samples. When noisy samples are prevalent, they may exhibit higher confidence than clean ones, leading to the prototype being biased toward the noisy samples. This biased prototype can cause many noisy samples to be misidentified as clean, thereby significantly degrading model performance in subsequent training stages.

Motivated by the observation that not all classes become ready for selection simultaneously in the presence of class imbalance and label noise, we shift the focus from how to select clean samples to when to select them. To this end, we propose a **R**eadiness-**A**ware **S**ample **S**election (RASS) strategy that identifies clean samples only from *ready-to-select* classes, where the model has sufficiently captured distinctive patterns, while delaying selection from *unready* classes until they become ready. To support learning distinctive features in unready classes, we also incorporate a novel negative learning scheme into RASS. As shown in Figure 1(b), RASS guides the model to focus on the clean samples even when initially biased to noisy ones during warm-up, leading to more reliable sample selection and improved generalization.

We summarize our main contributions as follows:

- We propose a novel sample selection strategy, RASS, that identifies clean samples only from *ready-to-select* classes where class-specific features are sufficiently learned, thereby avoiding premature sample selection in *unready* classes.
- To support the learning of discriminative features in unready classes, we incorporate a novel negative learning into RASS that enhances class separability using the samples selected as clean.
- Experimental results on benchmark and real-world datasets demonstrate that RASS outperforms existing methods by improving the quality of sample selection under noisy labeled datasets with class imbalance.

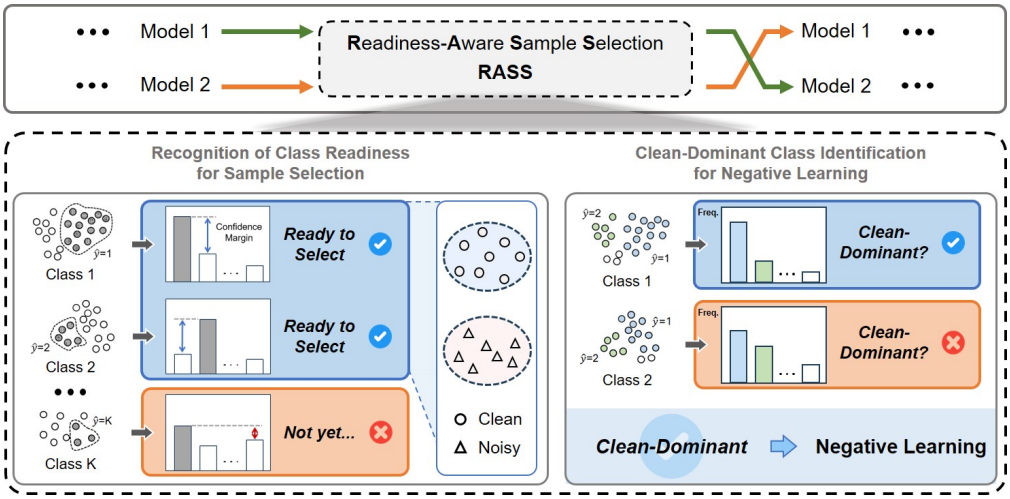


Figure 2: Schematic overview of the proposed method

2 Related Work

Learning with Noisy Labeled Data Since DNNs can easily overfit to noisy labeled data, numerous methods have been proposed to mitigate the impact of noisy labels. In particular, sample selection methods that treat small-loss instances as clean have attracted significant attention. Co-teaching [16] trains two models using small-loss samples selected by their peer model to reduce the selection bias. JoCoR [40] extended Co-teaching by introducing a joint loss function that encourages the predictions of the two models to be consistent. DivideMix [27] separates clean and noisy samples by fitting a Gaussian mixture model (GMM) to loss distributions, and incorporates a semi-supervised learning technique into the sample selection method. To address class imbalance in selected samples, Karim et al. [20] proposed a UNICON criterion that ensures class-balanced selection and improves selection performance by using contrastive learning. In other directions, various label correction and regularization methods have been explored to improve robustness against noisy labels [22, 30, 49, 7].

Learning with Class-Imbalanced Data Data imbalance has been typically addressed through re-weighting and re-sampling schemes. The re-weighting methods assign large importance to classes with low frequency [38, 17, 10]. On the other hand, the re-sampling methods balance class distributions by using over- or under-sampling techniques [14, 3, 36]. In a different direction from these studies, Cao et al. [4] proposed label-distribution-aware margin loss, which encourages the minority classes to have larger margins from the decision boundary. Similarly, Menon et al. [32] introduced a logit adjustment (LA) that promotes higher logit values for minority classes by using class prior information. While most of these methods are developed for fully supervised settings, several approaches have also been proposed for semi-supervised learning [19, 25, 50].

Learning with Noisy and Imbalanced Data Since most existing sample selection methods are implicitly designed for balanced datasets, several studies explored alternative strategies to identify clean samples in imbalanced cases. Wei et al. [41] proposed a prototypical noise detection method that leverages a distance-based metric between prototypes and each sample in the feature space. SFA [26] and DaSC [1] also employ class-specific prototypes for sample selection, while Huang et al. [18], Zhang et al. [47], and Fang et al. [12] account for class-dependent characteristics when identifying noisy samples. TABASCO [31] combines loss- and feature-based criteria in the proposed bi-dimensional selection strategy to improve the quality of selected samples. CBS [29] alleviates the class imbalance among selected samples by restricting the selection ratio for the majority classes. H2E [46] identifies hard noisy samples under class imbalance using a noise identifier that is iteratively refined to improve its robustness to imbalanced class distributions. Unlocker [45] introduces a plug-in framework that combines existing noisy label learning and long-tailed learning methods. In addition, Cao et al. [5] proposed to adjust the strength of Jacobian regularization based on the per-sample loss. Chen et al. [6] unified LA and ground-truth label estimation into a single framework.

3 Proposed Method

We consider an imbalanced and noisy labeled training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, where $\boldsymbol{x}_i \in \mathcal{X}$ and $y_i \in \{1, \dots, K\}$ are the input and annotated label of the i-th instance, respectively, and N is the total number of samples. The annotated label y_i can be different from the ground-truth label y_i^*. We denote the training samples annotated as class k by $\mathcal{D}_k = \{(\boldsymbol{x}_i, y_i) \mid y_i = k\}$, and the number of samples belonging to the k-th class as n_k. We assume that the classes are indexed in descending order of the sample size, i.e., $n_1 \geq \cdots \geq n_K$. $f : \mathcal{X} \to [0,1]^K$ denotes $K-$ary classification model. Given an input $\boldsymbol{x}_i$, the model produces a softmax probability vector $\boldsymbol{p}_i \in [0,1]^K$, where $\boldsymbol{p}_i[k]$ is the predicted probability of class k. The predicted label is denoted by $\hat{y} = \text{argmax}_k \boldsymbol{p}_i[k]$. For theoretical results, we defer rigorous statements, detailed assumptions, and complete proofs to the Appendix C.

3.1 Readiness-Aware Sample Selection

Figure 2 illustrates an overview of RASS. As in several sample selection methods [16, 27, 29], RASS is built on the co-teaching framework, where each model identifies clean samples and passes them to its peer. Unlike previous approaches that identify clean samples for all classes simultaneously, RASS selects clean ones only from the ready-to-select classes. RASS defers sample selection for the unready classes and facilitates their readiness through our novel negative learning scheme.

After a brief warm-up phase using the training dataset $\mathcal{D}$, RASS iterates the following steps: (1) recognition of class readiness to determine *ready-to-select* classes, (2) sample selection on ready-to-select classes, and (3) computation of negative loss to facilitate the readiness of unready classes. Detailed pseudo-codes are provided in Appendix B.

3.1.1 Recognition of Class Readiness

In RASS, we recognize ready-to-select classes by examining the predicted probability margins using samples predicted as the annotated class.

Specifically, for each class k ($k = 1, \dots, K$), we construct a matched set $\mathcal{M}_k$ using samples in $\mathcal{D}_k$ that are predicted as k. Then, we derive the probability vector $\bar{\boldsymbol{p}}_k = [\bar{p}_k^1, \dots, \bar{p}_k^K]$ by averaging the softmax probabilities of the samples in $\mathcal{M}_k$, where $\bar{p}_k^j$ denotes the averaged predicted probability for j-th class. If $\bar{p}_k^k$ satisfies Eq. (1), we consider that the distinctive patterns of class k have been sufficiently learned and the class is ready for selection.

$$\bar{p}_k^k > \alpha \times \bar{p}_k^{k'}, \tag{1}$$

where $k' = \text{argmax}_{j \neq k} \bar{p}_k^j$ and α (> 1) is a confidence margin. Here, we evaluate the margin between $\bar{p}_k^k$ and $\bar{p}_k^{k'}$. A small margin indicates that the model may struggle to differentiate class k from k', whereas a large margin implies that the model can effectively distinguish class k by capturing its distinctive features. Therefore, if we set α as a large enough value, a reliable sample selection can be performed for classes that satisfy Eq. (1), and we include such class indices in the ready-to-select class set $\mathcal{C}$.

Theorem 3.1 (Class Readiness)**.** *Let* $Z_i = p_i[k] - \alpha \max_{j \neq k} p_i[j]$ *and* $\Delta_k = \frac{1}{|\mathcal{M}_k|} \sum_{i \in \mathcal{M}_k} Z_i$. *Let* $\mathcal{E}_i = \mathbb{I}(Z_i \geq 0)$. *Let* $\omega_k = \mathbb{P}(\mathcal{E}_i | y_i^* = k, i \in \mathcal{M}_k) - \mathbb{P}(\mathcal{E}_i | y_i^* \neq k, i \in \mathcal{M}_k)$. *We assume* $\omega_k > 0$ *and* $\mathbb{P}(\mathcal{E}_i | y_i^* \neq k, i \in \mathcal{M}_k) \leq 0.5$. *Then, with probability at least* $1 - \delta$*,*

$$\text{Purity}(\mathcal{M}_k) \geq \omega_k^{-1} \left(\frac{\Delta_k + \alpha}{1 + \alpha} - \frac{1}{2} - \varepsilon_k(\delta) \right), \tag{2}$$

where $\varepsilon_k(\delta) = \sqrt{\log(2/\delta)/(2|\mathcal{M}_k|)}$.

The theorem states that Eq. (1) is a reasonable criterion for $\mathcal{M}_k$ to be mostly clean. When the average predicted probability for class k clearly exceeds the average for the second most probable by the factor α, then up to a small margin that shrinks with the number of samples in $\mathcal{M}_k$, most instances predicted as k are truly k. In practice, α controls how conservative the test is. Larger α demands greater separation and makes readiness more conservative, while having more data for class k makes the test more reliable because the error term $\varepsilon_k(\delta)$ scales like $|\mathcal{M}_k|^{-1/2}$. Consequently, if Eq. (1) holds with even a bit of extra margin, class k can be treated as ready and high-confidence samples can be safely selected from it. If the margin is small or negative, the class is not ready and should be

deferred. In particular, the condition $\bar{p}_k^k > \alpha \bar{p}_k^{k'}$ used in RASS can be viewed as a finite-sample proxy for ensuring that $\mathrm{Purity}(\mathcal{M}_k)$ is high under the assumptions of the theorem, whereas classes that fail this condition are deferred until the model has learned a more reliable separation. Our analysis of class readiness is provided in Appendix J.1.

3.1.2 Sample Selection

To derive a prototype for each ready-to-select class k in $\mathcal{C}$, we construct a representative sample set $\mathcal{R}_k$, which includes high-confidence samples that can serve as representative examples [31]. We begin by obtaining the maximum confidence score $\tilde{p}_k$, which is the largest value of $\boldsymbol{p}_i[k]$ among all samples $\boldsymbol{x}_i$ in $\mathcal{D}_k$. Then, we include the samples with a confidence score larger than $\tau \times \tilde{p}_k$ in $\mathcal{R}_k$, where τ $(0 \leq \tau < 1)$ is a confidence threshold. Using the samples in $\mathcal{R}_k$, the prototype $\boldsymbol{\mu}_k$ for class k is calculated as:

$$\boldsymbol{\mu}_k = \frac{1}{|\mathcal{R}_k|} \sum_{\boldsymbol{x}_i \in \mathcal{R}_k} \boldsymbol{z}_i, \tag{3}$$

where $\boldsymbol{z}_i$ is a feature vector of a sample $\boldsymbol{x}_i$ extracted from the penultimate layer of the model. Once the prototype is constructed, we compute the Euclidean distances $\{d_i\}_{i=1}^{N_k}$ for all samples in $\mathcal{D}_k$ from the prototype $\boldsymbol{\mu}_k$ in the feature space. Then, a two-component Gaussian mixture model is fitted to the distance distribution. The clean probability w_i of each sample $\boldsymbol{x}_i$ is defined as the posterior probability that the distance d_i belongs to the Gaussian component with the smaller mean.

Subsequently, samples with a clean probability greater than 0.5 are regarded as clean and included in a labeled set $\mathcal{S}_l$, while the remaining samples are treated as noisy and collected into an unlabeled set $\mathcal{S}_u$. For the unready classes that are not included in $\mathcal{C}$, we defer sample selection and include all samples in $\mathcal{S}_l$ without distinguishing clean from noisy samples. Formally, $\mathcal{S}_l$ and $\mathcal{S}_u$ are constructed as in Eqs. (4) and (5).

$$\mathcal{S}_l = \left(\bigcup_{k \in \mathcal{C}} \{(\boldsymbol{x}_i, y_i) \in \mathcal{D}_k \mid w_i > 0.5\} \right) \cup \left(\bigcup_{k \notin \mathcal{C}} \mathcal{D}_k \right). \tag{4}$$

$$\mathcal{S}_u = \bigcup_{k \in \mathcal{C}} \{(\boldsymbol{x}_i, y_i) \in \mathcal{D}_k \mid w_i \leq 0.5\}. \tag{5}$$

Below, we present our theoretical results on the purity improvement that can be achieved by RASS, as well as generalization bounds for models under the original noisy labeled dataset and the subset obtained by sample selection.

Theorem 3.2 (Purity Improvement)**.** *Let $k \in \mathcal{C}$ be a ready class. Let $\mathcal{D}_k$ be the observed set of samples with annotated label k and $\mathcal{D}_k^* \subset \mathcal{D}_k$ the subset with correct true label k. Denote $q_k = |\mathcal{D}_k^*|/|\mathcal{D}_k|$. Let $C_k = \{(\boldsymbol{x}_i, y_i) \in \mathcal{D}_k : w_i > 0.5\}$ be the subset selected by the prototype-based sample selection procedure described above. Assume that, in the feature space $\boldsymbol{z}_i$ extracted from the penultimate layer, clean and noisy samples are well separated (with a selection error bound $\gamma < 1/2$). Then, for $0 < q_k < 1$, the selected set C_k satisfies*

$$\mathrm{Purity}(C_k) > q_k \quad \textit{and} \quad \frac{|C_k|}{|\mathcal{D}_k|} \geq (1-\gamma) q_k. \tag{6}$$

Moreover, if the majority of the labels are clean ($q_k > 1/2$), then as $|\mathcal{D}_k| \to \infty$, we have $\mathrm{Purity}(C_k) \to 1$ and $|C_k|/|\mathcal{D}_k| \to q_k$.

This result implies that the sample selection improves the purity of the selected set relative to the noisy observed data. Moreover, the selection is nontrivial, retaining a decent proportion of the data rather than selecting only a few high-confidence samples. Asymptotically, the selected set becomes perfectly clean for classes satisfying the clean-majority condition. Therefore, for ready classes $k \in \mathcal{C}$, the selected subset C_k produced by RASS is intended to approximate the conditions of the theorem in practice, yielding both higher purity than the original $\mathcal{D}_k$ and a nonvanishing coverage of clean samples. We also examine the validity of the assumptions underlying Theorem 3.2 in Appendix D.

Lemma 3.3 (Excess Risk Bound under Noisy Labels)**.** *Let $\mathcal{D}$ be the clean data distribution and $\mathcal{F}$ a hypothesis class. Let the loss function ℓ be convex and bounded in $[0, L]$. Suppose the training set $S = S_c \cup S_n$ consists of N samples, where S_c contains N_c clean samples drawn i.i.d. from $\mathcal{D}$ and S_n contains $N_n = N - N_c$ samples with possibly noisy labels.*

Let $\hat{f} \in \mathcal{F}$ be the empirical risk minimizer on S:

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \ell(f(x_i), y_i), \tag{7}$$

and let $f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)]$ be the optimal predictor under the true distribution, with $R(f) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f(x), y)]$. Then, for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$R(\hat{f}) \leq R(f^*) + L\frac{N_n}{N_c} + 4L\mathfrak{R}_{N_c}(\mathcal{F}) + 2L\sqrt{\frac{\log(2/\delta)}{2N_c}}, \tag{8}$$

where $\mathfrak{R}_{N_c}(\mathcal{F})$ denotes the Rademacher complexity of $\mathcal{F}$ on N_c clean samples.

This generalization bound depends on two key factors: the number of clean samples, N_c, which controls the estimation error through the complexity and concentration terms, and the noise-to-clean ratio, N_n/N_c, which appears as an explicit noise penalty.

In contrast to the standard ERM bound in the clean-label setting, where all terms can vanish as $N \to \infty$, the noise penalty here does not disappear as long as N_n/N_c stays bounded away from zero. As a result, the bound does not certify that $R(\hat{f})$ converges to $R(f^*)$ under persistent label noise: it leaves a nonzero asymptotic gap driven by the noise-to-clean ratio. This viewpoint motivates a sample selection strategy: by constructing a training subset with higher label purity (a smaller number of noisy samples while retaining enough clean samples), we can tighten the bound by shrinking the dominant noise penalty term. The following theorem formalizes this trade-off.

Theorem 3.4 (Generalization Bound of Sample Selection)**.** *Let $B(N_c, N_n, \delta)$ be the generalization bound defined in Lemma 3.3. Let S_{sel} be a selected subset with N_c^{sel} clean and N_n^{sel} noisy samples, and let $N_{\text{sel}} = N_c^{\text{sel}} + N_n^{\text{sel}}$. With probability $\geq 1 - \delta$, $R(\hat{f}_{\text{select}}) \leq B(N_c^{\text{sel}}, N_n^{\text{sel}}, \delta)$. Suppose for large N_c, the selection strategy guarantees a strict noise reduction ($N_n^{\text{sel}}/N_{\text{sel}} \leq N_n/N - \varepsilon$ for some $\varepsilon > 0$) while retaining a fraction of clean data ($N_c^{\text{sel}} \geq cN_c$ for some $c \in (0, 1]$). If $\mathfrak{R}_m(\mathcal{F}) \to 0$ as $m \to \infty$, then for sufficiently large N_c,*

$$B(N_c^{\text{sel}}, N_n^{\text{sel}}, \delta) < B(N_c, N_n, \delta). \tag{9}$$

Moreover, if $N_c^{\text{sel}}/N_c \to 1$ and $N_n^{\text{sel}}/N_{\text{sel}} \to 0$ as $N_c \to \infty$ (i.e., the selection is consistent), then $B(N_c^{\text{sel}}, N_n^{\text{sel}}, \delta) \to B(N_c, 0, \delta)$, converging to the oracle bound.

3.1.3 Clean-Dominant Class Identification

In RASS, we leverage a novel negative learning that informs the model that the selected clean sample does not belong to its most confusing class. However, if the selected sample is a mislabeled one, our negative learning could be harmful, particularly when the most confusing class is the true class.

To reduce this potential harm, we restrict the negative learning to samples from clean-dominant classes, where clean instances are likely to dominate. We define a class k as clean-dominant if it belongs to $\mathcal{C}$, and the majority of predicted labels for samples in $\mathcal{D}_k$, denoted by $\hat{y}_k^{mode}$, is k. Since such classes can provide a relatively high proportion of clean instances, a more accurate sample selection can be performed compared to the other classes.

After identifying the clean-dominant classes from $\mathcal{C}$, we construct a sample set for negative learning $\mathcal{S}_n$. Specifically, $\mathcal{S}_n$ consists of samples in the labeled set $\mathcal{S}_l$ whose labels belong to the clean-dominant classes as in Eq. (10).

$$\mathcal{S}_n = \bigcup_{k \in \mathcal{C} \cap \{j | \hat{y}_j^{mode} = j\}} \{(\boldsymbol{x}_i, y_i) \in \mathcal{S}_l \mid y_i = k\} \tag{10}$$

3.1.4 Negative Loss

For each sample $\boldsymbol{x}_i$ in $\mathcal{S}_n$, the model assigns the most confusing class as a negative pseudo-label y_i^n as in Eq. (11).

$$y_i^n = \text{argmax}_{j \neq y_i} \boldsymbol{p}_i[j] \tag{11}$$

Table 1: Test accuracies on CIFAR-10 and CIFAR-100 datasets with various label noises and imbalance factors under the supervised objective. The best and second-best results are highlighted in **bold** and underlined, respectively.

Datasets	ρ	0.1		0.05		0.1		0.05		Avg
	Label Noise	Sym-40%	Sym-60%	Sym-40%	Sym-60%	Asym-20%	Asym-40%	Asym-20%	Asym-40%	Rank
CIFAR-10	CE	74.87±0.19	64.33±0.95	68.80±0.70	53.90±1.04	82.04±0.14	68.15±1.55	77.78±0.43	65.37±1.75	8.13**
	LDAM	75.81±0.30	62.52±0.34	66.27±2.20	52.58±1.42	83.33±0.30	69.38±0.79	79.56±0.93	64.55±0.23	8.13**
	LA	76.36±0.34	65.30±0.65	70.78±0.93	56.92±1.02	82.83±0.33	70.73±0.52	77.95±0.86	68.09±1.10	6.38**
	Co-teaching	73.70±0.26	62.97±0.56	66.07±0.70	48.78±1.26	78.78±0.21	66.23±0.14	69.95±3.06	58.78±5.45	10.63**
	JoCoR	81.72±0.70	43.67±1.35	72.95±2.31	40.85±2.74	84.50±0.29	58.31±4.77	81.14±0.27	56.03±8.45	8.50**
	DivideMix	81.98±0.22	69.57±0.28	77.93±0.22	58.93±0.45	84.58±0.23	71.50±1.07	81.63±0.23	70.03±1.60	4.00**
	HAR	73.90±1.39	63.81±0.27	67.25±1.23	52.40±2.01	81.43±0.51	70.49±1.14	77.87±0.70	66.79±2.57	8.38**
	RoLT	81.51±0.39	69.40±0.98	74.43±0.55	60.33±1.86	80.65±0.18	67.65±0.60	76.63±0.40	62.13±2.43	7.50**
	TABASCO	82.64±0.41	71.05±0.57	78.04±0.30	61.61±1.56	85.38±0.33	77.62±0.72	79.71±0.05	72.27±0.13	3.00**
	CBS	72.68±0.47	70.49±0.29	65.44±0.39	61.82±1.14	68.63±0.60	62.78±0.27	61.77±1.94	56.52±1.70	9.63**
	RobustLA	82.72±0.47	72.43±0.28	77.88±0.46	65.12±0.43	86.08±0.84	75.04±1.93	82.28±0.28	67.46±2.22	2.75**
	RASS	**83.35±0.44**	**73.64±0.31**	**78.96±0.26**	**66.64±0.39**	**86.36±0.05**	**81.23±0.40**	**83.73±0.34**	**77.53±0.56**	**1.00**
CIFAR-100	CE	38.74±0.71	26.58±0.52	32.29±0.13	22.65±0.45	49.76±0.50	37.16±0.27	44.82±0.25	32.32±0.36	10.63**
	LDAM	41.05±0.78	26.96±0.47	35.83±0.30	23.96±0.43	51.81±0.36	38.65±0.40	45.90±0.58	33.96±0.86	8.75**
	LA	40.05±0.81	27.08±0.63	34.47±0.26	23.53±0.46	50.92±0.13	40.20±0.33	46.37±0.32	36.05±0.33	8.38**
	Co-teaching	44.82±0.29	32.05±0.39	38.15±0.62	25.87±0.47	54.36±0.28	44.31±0.58	48.33±0.34	35.47±1.77	5.25**
	JoCoR	46.30±0.82	32.31±0.99	40.16±0.46	26.19±0.59	52.83±0.36	40.73±0.62	46.36±0.43	33.92±0.86	5.50*
	DivideMix	46.19±0.15	32.04±0.64	40.25±0.53	26.01±0.24	52.72±0.59	41.00±0.87	47.38±0.46	34.25±0.52	5.38*
	HAR	33.82±0.76	22.85±0.58	28.68±0.64	20.08±0.37	43.46±0.41	33.36±0.98	40.40±0.60	30.08±1.00	11.88**
	RoLT	44.58±0.67	30.41±0.18	38.08±0.31	26.28±0.17	54.50±0.74	41.70±0.62	48.24±0.25	36.64±0.60	5.00**
	TABASCO	46.26±0.31	32.39±0.21	**40.81±0.41**	26.18±0.50	**56.10±0.24**	48.32±0.53	**50.96±0.24**	41.42±0.45	2.38
	CBS	43.66±0.32	32.41±0.25	37.60±0.23	27.96±0.10	44.09±0.64	37.88±1.13	38.00±0.36	30.96±0.39	8.00**
	RobustLA	**46.81±0.18**	32.17±0.35	40.17±1.00	27.87±0.83	51.61±0.73	40.99±0.47	46.61±0.61	36.00±0.34	4.63
	RASS	45.77±0.03	**32.98±0.49**	39.60±0.05	**28.41±0.50**	55.83±0.13	**48.92±0.19**	50.11±0.86	**42.05±0.35**	**2.25**

$$\mathcal{L}_n = -\frac{1}{|\mathcal{S}_n|} \sum_{i \in \mathcal{S}_n} \log\left(1 - \boldsymbol{p}_i[y_i^n]\right). \tag{12}$$

Then, the negative loss in Eq. (12) encourages the model to suppress the predicted probability for y_i^n-th class to zero, which helps to mitigate confusion. While our negative loss primarily reduces confusion for samples in clean-dominant classes, it also brings benefits to other classes. As the confusion of samples in $\mathcal{S}_n$ reduces, the model learns to distinguish clean-dominant classes from other classes. From the perspective of those other classes, noisy samples, whose true labels belong to clean-dominant classes, can also be pushed away from the clean samples. As a result, our negative learning scheme helps unready classes learn discriminative features, enabling those classes to reach readiness for sample selection earlier.

3.2 Training Objectives

RASS can be plugged into both supervised and semi-supervised training objectives. In the supervised objective, only the samples regarded as labeled are used for training, whereas the semi-supervised objective leverages both labeled and unlabeled samples obtained by RASS. Since our negative loss is derived from a part of the samples that are regarded as labeled, it is utilized in both objectives.

3.2.1 Supervised Objective

Here, we train the model with loss $\mathcal{L}_{sup}$ as follows:

$$\mathcal{L}_{sup} = \mathcal{L}_{ce} + \lambda_1 \cdot \mathcal{L}_n + \lambda_2 \cdot \mathcal{L}_{reg}, \tag{13}$$

where $\mathcal{L}_{ce}$ and $\mathcal{L}_{reg}$ denote the cross-entropy and the KL-divergence loss, respectively, both computed using the labeled set $\mathcal{S}_l$. As in previous works, $\mathcal{L}_{reg}$ is employed to regularize the predictions toward a uniform distribution to prevent overfitting against noisy labels [27, 31]. $\mathcal{L}_n$ denotes our negative loss, computed from the sample set $\mathcal{S}_n$. The terms λ_1 and λ_2 are weights that balance loss functions.

3.2.2 Semi-Supervised Objective

Previous sample selection studies under semi-supervised learning commonly adopted MixMatch as a standard training framework [27, 31]. Following this convention, we train the model using the loss $\mathcal{L}_{semi}$ as follows:

$$\mathcal{L}_{semi} = \mathcal{L}_{MixMatch} + \lambda_1 \cdot \mathcal{L}_n + \lambda_2 \cdot \mathcal{L}_{reg}, \tag{14}$$

where $\mathcal{L}_{MixMatch}$ is the MixMatch loss [2], which leverages $\mathcal{S}_l$ and $\mathcal{S}_u$ as labeled and unlabeled data, respectively. In this objective, $\mathcal{L}_{reg}$ is computed using the entire samples in the dataset $\mathcal{D}$.

Table 2: Test accuracies on CIFAR-10 and CIFAR-100 datasets with various label noises and imbalance factors under the semi-supervised objective. The best and second-best results are highlighted in **bold** and underlined, respectively.

Datasets	ρ	0.1		0.05		0.1		0.05		Avg
	Label Noise	Sym-40%	Sym-60%	Sym-40%	Sym-60%	Asym-20%	Asym-40%	Asym-20%	Asym-40%	Rank
CIFAR-10	Co-teaching	71.40±2.60	45.85±1.81	52.22±0.66	41.32±2.24	82.95±0.86	59.16±4.00	72.09±1.20	53.03±5.71	5.63**
	JoCoR	76.51±0.72	41.83±1.55	61.24±5.00	39.05±1.65	81.60±4.50	57.36±4.26	72.55±2.74	31.54±0.72	6.00**
	DivideMix	83.33±0.14	81.48±0.30	77.64±0.50	71.73±4.37	82.25±0.13	79.12±3.40	75.36±1.20	71.84±1.75	3.25**
	RoLT	79.85±0.29	79.70±0.37	77.04±0.49	75.90±0.86	77.58±0.14	68.04±1.53	69.04±1.56	64.95±1.00	4.38**
	TABASCO	84.17±0.75	83.81±0.35	81.22±0.49	78.05±0.43	82.56±0.38	80.84±1.28	77.72±0.42	74.61±0.40	2.13**
	CBS	75.68±0.36	74.53±0.31	70.26±0.29	66.86±0.91	70.92±0.30	66.27±1.02	68.12±0.28	61.95±1.58	5.63**
	RASS	**84.50±0.16**	**83.89±0.08**	**81.87±0.10**	**78.18±0.66**	**83.28±0.62**	**85.53±0.33**	**78.27±0.05**	**76.24±0.64**	**1.00**
CIFAR-100	Co-teaching	51.59±0.46	36.17±0.28	43.39±0.28	31.86±1.47	55.52±0.45	39.33±0.46	47.01±0.41	32.23±0.83	5.63**
	JoCoR	48.75±0.87	37.04±0.31	41.99±1.11	31.54±0.72	52.67±0.73	36.42±0.47	44.92±0.42	30.03±0.31	6.63**
	DivideMix	**55.88±0.12**	46.09±0.55	49.49±0.19	38.38±0.81	58.85±0.21	45.38±0.36	51.70±0.08	38.18±0.32	3.13
	RoLT	54.71±0.68	46.86±0.32	49.42±0.38	39.67±0.45	56.03±0.45	43.57±0.37	49.36±0.47	40.40±0.84	3.50**
	TABASCO	55.39±0.30	46.73±0.63	**50.10±0.12**	39.95±0.38	**59.62±0.17**	51.07±0.47	**53.84±0.38**	44.97±0.04	1.88
	CBS	49.27±0.58	38.44±0.41	42.57±0.25	33.26±0.92	44.09±0.64	42.52±0.52	38.60±0.57	39.99±0.63	5.63**
	RASS	55.48±0.06	**46.89±0.12**	49.91±0.25	**40.40±0.53**	58.43±0.45	**52.31±0.08**	53.37±0.30	**45.22±0.91**	**1.63**

Note that we retain all samples from unready classes in $\mathcal{S}_l$ to preserve training signals for those classes under both objectives while avoiding unreliable pseudo-labels in semi-supervised training.

4 Experiments

4.1 Experimental Setup

Details about datasets and implementations are provided in Appendix E.

Datasets We validate RASS on two benchmark datasets, CIFAR-10 and CIFAR-100, as well as on two real-world datasets, Clothing1M and Animal-10N. Following prior works [31, 48], we first simulated class imbalance and then injected synthetic label noise into the benchmark datasets. To simulate class imbalance, we decreased the number of instances in each class exponentially, based on the imbalance factor $\rho = n_K/n_1$. To corrupt the imbalanced dataset, we injected symmetric (Sym), asymmetric (Asym), or instance-dependent (IDN) label noise as much as the noise ratio. For the real-world datasets with inherent label noise, we simulated class imbalance as we described above. We also evaluate on large-scale datasets and more realistic class-imbalanced scenarios in Appendix G.

Comparison Methods To demonstrate the effectiveness of the proposed method, we employ three types of baselines as follows: (1) Class-imbalance learning methods, such as LDAM [4] and LA [32]; (2) Methods for learning with noisy labeled data, including Co-teaching [16], JoCoR [40], and DivideMix [27]; (3) Methods that address both class imbalance and label noise simultaneously, such as HAR [5], RoLT [41], TABASCO [31], CBS [29], and RobustLA [6]. Additional comparisons with H2E [46], SFA [26], and SSBL [47] are provided in Appendix F.

Implementation For CIFAR-10/100, we trained 100 epochs using PreAct ResNet-18. For Clothing1M, we trained 80 epochs using ResNet-50, and for Animal-10N, we trained 100 epochs using VGG-19. Under the supervised objective, we evaluated all baselines. Under the semi-supervised objective, we only evaluated baselines that involve sample selection methods [16, 40, 27, 41, 31, 29], as these methods are designed to selectively utilize training samples and are suited for semi-supervised learning. In all experiments, we set the confidence margin α to 3 and the confidence threshold τ to 0.8. For the supervised objective, the loss weights λ_1 and λ_2 were set to 1e-3 and 1, respectively. For the semi-supervised objective, the loss weights λ_1 and λ_2 were set to 1e-1 and 1, respectively. Results about the sensitivity analysis on these hyperparameters are given in Appendix H.1.

4.2 Experimental Results

4.2.1 Results on Benchmark Datasets

We evaluated the classification performance and average rank of RASS on the benchmark datasets. Tables 1 and 2 present the results under Sym and Asym label noise for the supervised and semi-supervised objectives, respectively. Additional results under IDN noise are provided in Appendix I. Here, CE refers to the baseline model trained with the cross-entropy loss. Superscripts * and ** indicate that the p-value of the Wilcoxon signed-rank test is less than 0.05 and 0.01, respectively.

Table 3: Test accuracies on the real-world datasets with different imbalance factors. The best and second-best results are highlighted in **bold** and underlined, respectively.

Obj.	Dataset	ρ	CE	LDAM	LA	Co-teaching	JoCoR	DivideMix	HAR	RoLT	TABASCO	CBS	Robust LA	RASS
Supervised	Clothing1M	0.1	49.07 ± 0.48	55.46 ± 0.95	53.55 ± 0.01	29.60 ± 1.47	38.23 ± 0.86	65.07 ± 1.47	44.44 ± 0.72	65.29 ± 0.45	51.37 ± 0.79	34.70 ± 0.84	54.66 ± 1.74	**66.27** ± 0.76
		0.05	41.06 ± 1.22	47.31 ± 1.00	47.67 ± 0.25	24.60 ± 0.23	32.46 ± 0.08	59.28 ± 0.78	41.50 ± 0.63	59.64 ± 0.47	43.21 ± 1.86	28.22 ± 0.68	42.38 ± 1.70	**60.14** ± 1.04
	Animal-10N	0.1	70.42 ± 0.46	77.32 ± 0.04	74.97 ± 0.40	71.00 ± 0.43	75.41 ± 0.11	76.08 ± 0.26	70.26 ± 1.16	77.38 ± 0.28	76.91 ± 0.32	56.77 ± 0.21	77.41 ± 0.28	**77.72** ± 0.20
		0.05	65.98 ± 1.01	74.44 ± 0.37	72.76 ± 0.69	67.40 ± 0.45	72.70 ± 0.34	73.50 ± 0.37	64.55 ± 0.45	74.86 ± 0.24	74.82 ± 0.36	55.41 ± 0.36	74.18 ± 0.11	**75.16** ± 0.22
Semi-supervised	Clothing1M	0.1	-	-	-	37.11 ± 3.71	39.25 ± 1.15	61.51 ± 3.64	-	62.92 ± 1.39	47.67 ± 1.12	42.31 ± 1.62	-	**63.91** ± 1.01
		0.05	-	-	-	31.64 ± 1.63	33.62 ± 3.98	54.01 ± 2.58	-	55.68 ± 1.00	40.88 ± 0.83	36.14 ± 0.65	-	**55.95** ± 1.07
	Animal-10N	0.1	-	-	-	75.63 ± 0.53	75.26 ± 0.50	76.57 ± 0.49	-	76.60 ± 0.24	76.77 ± 0.25	60.44 ± 0.52	-	**77.22** ± 0.49
		0.05	-	-	-	72.06 ± 0.21	72.45 ± 0.59	74.52 ± 0.23	-	74.88 ± 0.16	74.26 ± 0.15	58.20 ± 0.21	-	**74.92** ± 0.11

As presented in Table 1, RASS outperformed comparison methods in most cases. For the CIFAR-10, RASS achieved the best performance in all cases. Notably, for the case of Asym-40% noise under the imbalance factors 0.1 and 0.05, RASS improved the second-best accuracy by 3.61% and 5.26%, respectively. For the CIFAR-100, RASS consistently ranked in the top two in most cases, and achieved the best performance in highly corrupted datasets, such as Sym-60% and Asym-40%, regardless of the imbalance factors. Class-imbalance learning methods, such as LDAM and LA, exhibited suboptimal performance because they do not consider the presence of noisy labels. For the approaches that deal with noisy labels, such as Co-teaching, JoCoR, and DivideMix, they showed competitive results when the imbalance factor was set to 0.1. However, their effectiveness decreased under severe class imbalance, where the imbalance factor was decreased to 0.05. Although TABASCO and RobustLA, which are designed to address both noisy labels and class imbalance, showed decent performance across various label noises and imbalance factors, their results were still inferior to those of RASS. The results for head and tail classes are reported in Appendix J.2.

As shown in Table 2, under the semi-supervised objective, RASS also outperformed the comparison methods in most cases, achieving the best average rank on both CIFAR-10 and CIFAR-100. For CIFAR-10, the proposed method achieved the best performance in all cases. Similar to the supervised results, RASS significantly improved the second-best method under asymmetric noise cases, where labels are flipped only between pre-defined class pairs, making it hard to distinguish clean samples from noisy ones. For CIFAR-100, RASS achieved the best results under the Sym-60% and Asym-40% cases, showing strong robustness to severe label noise under the class-imbalanced datasets. These results highlight the effectiveness of RASS in both supervised and semi-supervised objectives.

Despite these performance gains, the improvements achieved by RASS are relatively small under low-noise conditions on CIFAR-100. When the noise level is low, early feature representations can be more reliable, reducing the benefit of delaying sample selection. In addition, CIFAR-100 contains many classes with relatively few samples per class, which can make class-wise prototypes less representative. RASS may therefore provide less benefit when delaying sample selection offers little advantage or when class-wise prototypes are less reliable.

4.2.2 Results on Real-World Datasets

We validated the proposed method on the two real-world noisy labeled datasets, Clothing1M and Animal-10N. Table 3 shows the classification accuracy under the supervised and semi-supervised objectives. On both datasets, RASS showed the best performance across all imbalance factors regardless of training objectives. Notably, although RoLT showed second-best results in most cases in Clothing1M and Animal-10N, it was less effective on the benchmark datasets. For TABASCO and RobustLA, their effectiveness was limited to the benchmark datasets. In contrast, RASS consistently outperformed all baselines on both benchmark and real-world datasets, which highlights the broad applicability of RASS to various datasets.

4.2.3 Ablation Study

We conducted an ablation study using the CIFAR-10 dataset with Sym-60% and Asym-40% under different imbalance factors. The ablation models include: (1) *w/o RA*: identifying clean samples without employing our readiness-aware strategy, (2) *w/o NL*: training the model without negative learning, and (3) *w/o PS*: identifying clean samples based on the small-loss criterion instead of the prototype-based approach.

Table 4: Ablation results on the CIFAR-10 dataset with various label noises and imbalance factors.

Obj.	Label Noise	Sym-60%		Asym-40%	
	ρ	0.1	0.05	0.1	0.05
Sup	RASS	73.64	66.64	81.23	77.53
	w/o RA	72.49 (-1.15)	64.15 (-2.49)	79.56 (-1.67)	69.30 (-8.23)
	w/o NL	73.56 (-0.08)	66.24 (-0.40)	79.94 (-1.29)	77.08 (-0.45)
	w/o PS	69.47 (-4.17)	62.44 (-4.20)	72.75 (-8.48)	68.29 (-9.24)
Semi	RASS	83.89	78.18	85.53	76.24
	w/o RA	82.71 (-1.18)	77.16 (-1.02)	74.25 (-11.28)	72.09 (-4.15)
	w/o NL	83.03 (-0.86)	77.29 (-0.89)	84.87 (-0.66)	75.84 (-0.40)
	w/o PS	82.07 (-1.82)	71.82 (-6.36)	80.92 (-4.61)	72.01 (-4.23)

As shown in Table 4, *w/o RA* resulted in a significant performance drop. The degradation was remarkable under asymmetric noise, where labels are flipped only between pre-defined class pairs, making it challenging to distinguish clean samples from noisy ones. Specifically, under the supervised objective with Asym-40% noise, classification accuracy decreased by 1.67% and 8.23% when the imbalance factor ρ was set to 0.1 and 0.05, respectively. In the semi-supervised learning setting, the degradation was 11.28% and 4.15% for each case. These results demonstrate the effectiveness of the proposed sample selection strategy that identifies clean samples only for the ready classes. *w/o NL* also consistently led to a performance drop in all cases. Although the impact was relatively less than *w/o RA*, we confirmed that negative learning still helps RASS achieve improved performance. *w/o PS* caused a considerable performance drop in all variants. While the proposed readiness-aware strategy and negative learning can boost the generalization ability, we observed that the effectiveness of RASS is limited when it relies on the small-loss criterion.

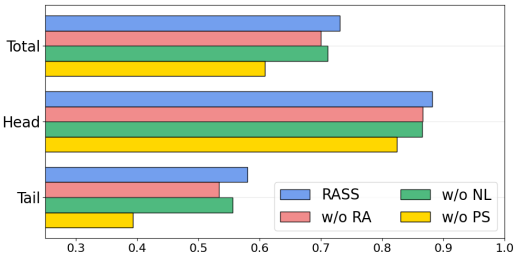


Figure 3: F1-score of sample selection results from the Total, Head, and Tail classes on the CIFAR-10 with Sym-60% label noise when the imbalance factor is set to 0.05. *Total* denotes all classes. *Head* and *Tail* refer to half of the classes with a large number of samples and the others, respectively.

In addition, under the supervised objective, we analyzed the effect of each component on the quality of selected samples using the CIFAR-10 with Sym-60% noise, when the imbalance factor was set to 0.05. We compared the F1-score to account for both the purity and coverage of selected samples. Here, *purity* is the proportion of clean samples among the selected ones, and *coverage* is the proportion of selected clean samples among all clean ones. As shown in Figure 3, RASS showed superiority in selecting clean samples from noisy labeled and class-imbalanced datasets. When either the RA or NL was removed, the F1-score decreased, especially for the Tail classes. Further ablation results and sample selection results on CIFAR-100 are provided in Appendices H.2 and K, respectively.

5 Conclusion

We introduced RASS, a readiness-aware sample selection strategy for noisy labeled datasets with class imbalance. In RASS, sample selection is allowed only for ready-to-select classes, thereby avoiding premature selection in unready classes. To facilitate the readiness of unready classes by enhancing class separability, we incorporated a novel negative learning scheme into RASS. Experiments on both benchmark and real-world datasets demonstrated that RASS improves generalization performance and quality of selected samples for the datasets with diverse label noise and class imbalance. For future work, we can extend our method to handle noisy labeled datasets in more practical scenarios, such as domain shift or open-set conditions, which can commonly arise in real-world applications. Another direction is to explore class-adaptive readiness criteria with class-specific values of α to account for different learning dynamics across classes.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (RS-2026-25603088 and RS-2026-25597532); by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2025-25420424); by the Korea Meteorological Administration Research and Development Program (KMI; RS-2025-02221093 and RS-2025-02219688); and by the Advanced GPU Utilization Support Program funded by MSIT (Prof. Hoki Kim, Chung-Ang University).

References

[1] J. S. Baik, I. Y. Yoon, K. H. Kim, and J. W. Choi. Distribution-aware robust learning from long-tailed data with noisy labels. In *European Conference on Computer Vision*, pages 160–177. Springer, 2024.

[2] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

[3] J. Byrd and Z. Lipton. What is the effect of importance weighting in deep learning? In *International conference on machine learning*, pages 872–881. PMLR, 2019.

[4] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

[5] K. Cao, Y. Chen, J. Lu, N. Arechiga, A. Gaidon, and T. Ma. Heteroskedastic and imbalanced deep learning with adaptive regularization. *arXiv preprint arXiv:2006.15766*, 2020.

[6] M. Chen, Y. Du, W. Jiang, B. Zhang, S. Feng, Y. Xin, and C. Wang. Robust logit adjustment for learning with long-tailed noisy data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 15830–15838, 2025.

[7] C. Choi, W. Lee, and Y. Son. Multi-stage ensemble with refinement for noisy labeled data classification. *Expert Systems with Applications*, 255:124672, 2024.

[8] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

[9] W. Dai, P. Chen, M. Lu, D. Li, H. Wei, H. Cui, and P. P. Liang. Climb: Data foundations for large scale multimodal clinical foundation models. *arXiv preprint arXiv:2503.07667*, 2025.

[10] W. Dang, Z. Yang, W. Dong, X. Li, and G. Shi. Inverse weight-balancing for deep long-tailed learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11713–11721, 2024.

[11] M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. P. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International conference on machine learning*, pages 7480–7512. PMLR, 2023.

[12] C. Fang, L. Cheng, Y. Mao, D. Zhang, Y. Fang, G. Li, H. Qi, and L. Jiao. Separating noisy samples from tail classes for long-tailed image classification with label noise. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):16036–16048, 2023.

[13] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from internet image searches. *Proceedings of the IEEE*, 98(8):1453–1466, 2010.

[14] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla. Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *Journal of artificial intelligence research*, 61:863–905, 2018.

[15] T. Fredriksson, D. I. Mattos, J. Bosch, and H. H. Olsson. Data labeling: An empirical investigation into industrial challenges and mitigation strategies. In *International conference on product-focused software process improvement*, pages 202–216. Springer, 2020.

[16] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

[17] C. Huang, Y. Li, C. C. Loy, and X. Tang. Deep imbalanced learning for face recognition and attribute prediction. *IEEE transactions on pattern analysis and machine intelligence*, 42(11): 2781–2794, 2019.

[18] Y. Huang, B. Bai, S. Zhao, K. Bai, and F. Wang. Uncertainty-aware learning against label noise on imbalanced datasets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 6960–6969, 2022.

[19] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

[20] N. Karim, M. N. Rizve, N. Rahnavard, A. Mian, and M. Shah. Unicon: Combating label noise through uniform selection and contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9676–9686, 2022.

[21] T. Kim, J. Ko, J. Choi, S.-Y. Yun, et al. Fine samples for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:24137–24149, 2021.

[22] Y. Kim, J. Yim, J. Yun, and J. Kim. Nlnl: Negative learning for noisy labels. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 101–110. IEEE, 2019.

[23] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images, 2009.

[24] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[25] H. Lee, S. Shin, and H. Kim. Abc: Auxiliary balanced classifier for class-imbalanced semi-supervised learning. *Advances in Neural Information Processing Systems*, 34:7082–7094, 2021.

[26] H.-T. Li, T. Wei, H. Yang, K. Hu, C. Peng, L.-B. Sun, X.-L. Cai, and M.-L. Zhang. Stochastic feature averaging for learning with long-tailed noisy labels. In *IJCAI*, pages 3902–3910, 2023.

[27] J. Li, R. Socher, and S. C. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020.

[28] Z. Lian, H. Chen, L. Chen, H. Sun, L. Sun, Y. Ren, Z. Cheng, B. Liu, R. Liu, X. Peng, et al. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. *arXiv preprint arXiv:2501.16566*, 2025.

[29] H. Liu, M. Sheng, Z. Sun, Y. Yao, X.-S. Hua, and H.-T. Shen. Learning with imbalanced noisy data by preventing bias in sample selection. *IEEE Transactions on Multimedia*, 2024.

[30] S. Liu, J. Niles-Weed, N. Razavian, and C. Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33: 20331–20342, 2020.

[31] Y. Lu, Y. Zhang, B. Han, Y.-m. Cheung, and H. Wang. Label-noise learning with intrinsically long-tailed data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1369–1378, 2023.

[32] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

[33] B. Pan, Q. Li, X. Tang, W. Huang, Z. Fang, F. Liu, J. Wang, J. Yu, and Y. Shi. Nlprompt: Noise-label prompt learning for vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19963–19973, 2025.

[34] H. Song, M. Kim, and J.-G. Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *International conference on machine learning*, pages 5907–5915. PMLR, 2019.

[35] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34 (11):8135–8153, 2022.

[36] H. Song, M. Kim, and J.-G. Lee. Toward robustness in multi-label classification: A data augmentation strategy against imbalance and noise. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 21592–21601, 2024.

[37] G. Van Horn and P. Perona. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017.

[38] Y.-X. Wang, D. Ramanan, and M. Hebert. Learning to model the tail. *Advances in neural information processing systems*, 30, 2017.

[39] C. Wei, K. Sohn, C. Mellina, A. Yuille, and F. Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10857–10866, 2021.

[40] H. Wei, L. Feng, X. Chen, and B. An. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13726–13735, 2020.

[41] T. Wei, J.-X. Shi, W.-W. Tu, and Y.-F. Li. Robust long-tailed learning under label noise. *arXiv preprint arXiv:2108.11569*, 2021.

[42] M. Xia, Z. Huang, R. Wu, G. Lyu, J. Zhao, G. Chen, and H. Wang. Unbiased multi-label learning from crowdsourced annotations. In *Forty-first International Conference on Machine Learning*, 2024.

[43] X. Xia, T. Liu, B. Han, N. Wang, M. Gong, H. Liu, G. Niu, D. Tao, and M. Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. *Advances in neural information processing systems*, 33:7597–7610, 2020.

[44] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015.

[45] H. Xu, R. Zhang, M. Li, Y. Zhang, Y. Lu, B. Han, Y.-m. Cheung, H. Wang, et al. Unlocker: Disentangle the deadlock of learning between label-noisy and long-tailed data. *Advances in Neural Information Processing Systems*, 38:130302–130329, 2026.

[46] X. Yi, K. Tang, X.-S. Hua, J.-H. Lim, and H. Zhang. Identifying hard noise in long-tailed sample distribution. In *European Conference on Computer Vision*, pages 739–756. Springer, 2022.

[47] L. Zhang, Z.-H. Tian, W. Zhou, and W. Wang. Learning from long-tailed noisy data with sample selection and balanced loss. In *IJCAI*, pages 5471–5480, 2024.

[48] M. Zhang, X. Zhao, J. Yao, C. Yuan, and W. Huang. When noisy labels meet long tail dilemmas: A representation calibration method. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15890–15900, 2023.

[49] G. Zheng, A. H. Awadallah, and S. Dumais. Meta label correction for noisy label learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11053–11061, 2021.

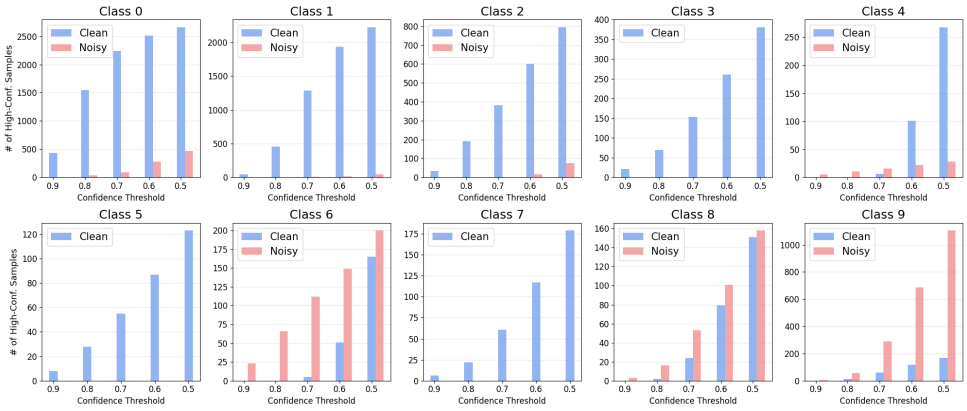
[50] H. Zheng, L. Zhou, H. Li, J. Su, X. Wei, and X. Xu. Bem: Balanced and entropy-based mix for long-tailed semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22893–22903, 2024.

A Complete Results for Figure 1

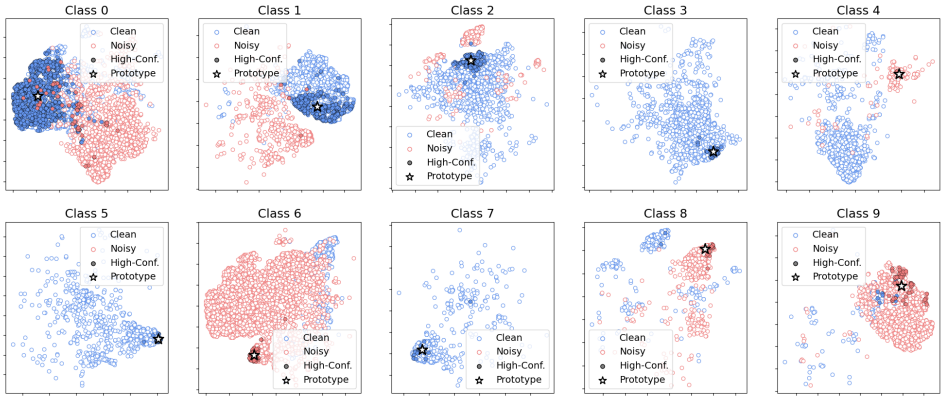
In Figure 1 of the main text, we illustrated a potential failure case of existing sample selection methods and demonstrated the effectiveness of RASS. This example is based on the CIFAR-10 with 40% asymmetric label noise and an imbalance factor of 0.1, focusing on class 6. To provide a more comprehensive view, in Figures A1 and A2, we presented per-class visualizations after the warm-up phase and after 10 training epochs with RASS, respectively.

As shown in Figure A1, for classes containing a sufficient number of clean samples, the model successfully captures their distinctive features, assigning higher confidence scores to clean instances. However, for classes with fewer clean samples, such as classes 4, 6, 8, and 9, the model struggles to learn meaningful patterns. In these cases, noisy samples often exhibit higher confidence scores than clean ones, leading to biased class prototypes toward noisy data. This makes it difficult for existing prototype-based sample selection methods to accurately identify clean instances.

In contrast, as shown in Figure A2, RASS effectively guides the model to focus on clean samples, even in classes where the model initially failed to distinguish clean from noisy data. This further supports the effectiveness of RASS, which identifies clean instances only from *ready-to-select* classes while promoting the readiness of remaining *unready* classes through our novel negative learning mechanism.

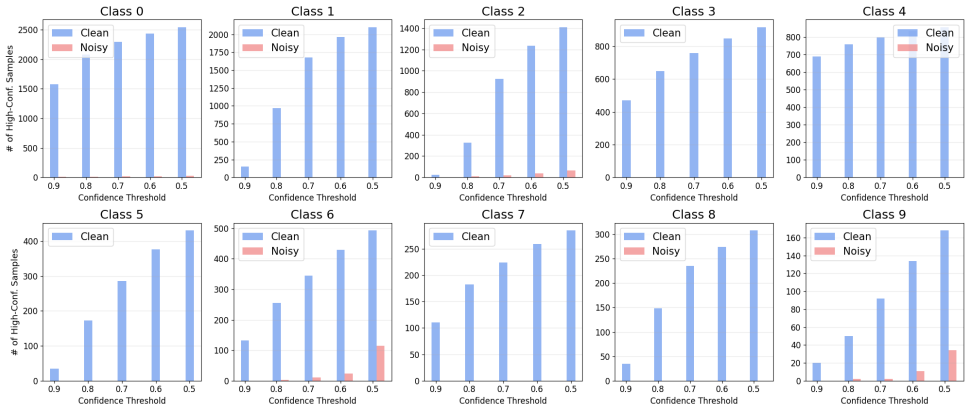


(a) The number of clean and noisy instances among high-confidence samples

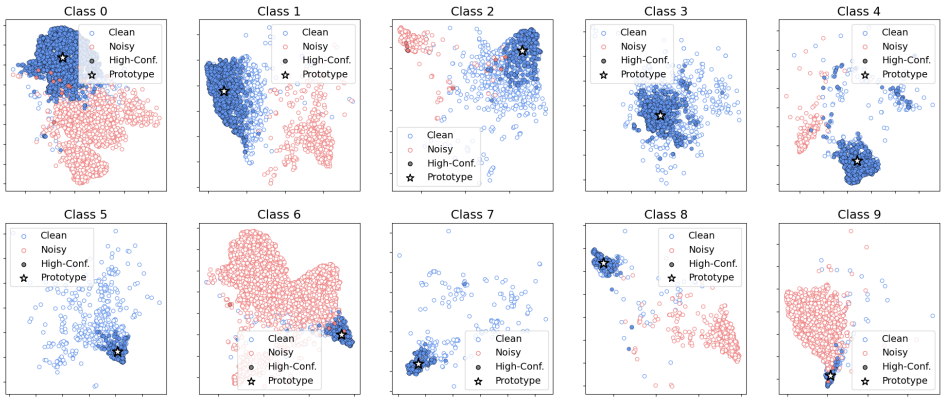


(b) Feature representations of clean and noisy samples, along with the prototype

Figure A1: Visualization of (a) the composition of high-confidence samples and (b) the feature representations after warm-up on CIFAR-10 with Asym-40% label noise when the imbalance factor is 0.1.



(a) The number of clean and noisy instances among high-confidence samples



(b) Feature representations of clean and noisy samples, along with the prototype

Figure A2: Visualization of (a) the composition of high-confidence samples and (b) the feature representations after 10 epochs of training with RASS on CIFAR-10 with Asym-40% label noise when the imbalance factor is 0.1.

B Pseudo-Codes

In Algorithm B1, we present the overall training process using RASS. Also, we describe the procedure for obtaining the ready-to-select class set, the labeled and unlabeled sample sets, and the sample set for negative learning in Algorithms B2, B3, and B4. Please refer to the main text for the definitions of all symbols and equations.

Algorithm B1 Overall Training Process

Input: dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, model f^1 and f^2
Parameter: confidence margin $(\alpha > 1)$, confidence threshold $(0 \leq \tau < 1)$, warm-up epochs E_w, max epochs E_{max}
Output: trained models f^1 and f^2

warm-up f^1 and f^2 using $\mathcal{D}$ for E_w epochs
for epoch $e = E_w + 1, \ldots, E_{max}$ **do**
  for model index $m \in \{1, 2\}$ **do**
    $\mathcal{C}^m \leftarrow$ Recognition of Class Readiness using $\mathcal{D}$ and f^m (Algorithm B2)
    $\mathcal{S}_l^m, \mathcal{S}_u^m \leftarrow$ Sample Selection using $\mathcal{D}$, $\mathcal{C}^m$, and f^m (Algorithm B3)
    $\mathcal{S}_n^m \leftarrow$ Clean-Dominant Class Identification using $\mathcal{D}$, $\mathcal{C}^m$, $\mathcal{S}_l^m$, and f^m (Algorithm B4)
  end for
  if supervised objective **then**
    train f^1 and f^2 using $\mathcal{S}_l^2$, $\mathcal{S}_n^2$ and $\mathcal{S}_l^1$, $\mathcal{S}_n^1$, respectively, with $\mathcal{L}_{sup}$ as defined in Eq. (13)
  end if
  if semi-supervised objective **then**
    train f^1 and f^2 using $\mathcal{S}_l^2$, $\mathcal{S}_n^2$ and $\mathcal{S}_l^1$, $\mathcal{S}_n^1$, respectively, with $\mathcal{L}_{semi}$ as defined in Eq. (14)
  end if
end for

Algorithm B2 Recognition of Class Readiness

Input: dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, and model f
Parameter: confidence margin $(\alpha > 1)$
Output: ready-to-select class set $\mathcal{C}$

$\mathcal{C} \leftarrow \emptyset$
for class $k = 1, \cdots, K$ **do**
  fetch class-wise training dataset $\mathcal{D}_k$
  $\{\boldsymbol{p}_i\}_{i=1}^{|\mathcal{D}_k|}, \leftarrow f(\boldsymbol{x}_i)$ for $\boldsymbol{x}_i \in \mathcal{D}_k$
  $\mathcal{M}_k \leftarrow \{\boldsymbol{x}_i \mid (\boldsymbol{x}_i, y_i) \in \mathcal{D}_k, \hat{y}_i = k\}$
  $\bar{\boldsymbol{p}}_k \leftarrow \frac{1}{|\mathcal{M}_k|} \sum_{\boldsymbol{x}_i \in \mathcal{M}_k} \boldsymbol{p}_i$
  $\bar{p}_k^k \leftarrow k$-th element of $\bar{\boldsymbol{p}}_k$
  if $\bar{p}_k^k$ satisfies Eq. (1) **then**
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{k\}$
  end if
end for

Algorithm B3 Sample Selection

Input: dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, ready-to-select class set $\mathcal{C}$, and model f
Parameter: confidence threshold $(0 \leq \tau < 1)$
Output: labeled sample set $\mathcal{S}_l$, unlabeled sample set $\mathcal{S}_u$

$\mathcal{S}_l, \mathcal{S}_u \leftarrow \emptyset$
for class $k = 1, \cdots, K$ **do**
  fetch class-wise training dataset $\mathcal{D}_k$
  if $k \in \mathcal{C}$ **then**
    $\{\boldsymbol{p}_i\}_{i=1}^{|\mathcal{D}_k|}, \{\boldsymbol{z}_i\}_{i=1}^{|\mathcal{D}_k|} \leftarrow f(\boldsymbol{x}_i)$ for $\boldsymbol{x}_i \in \mathcal{D}_k$
    $\tilde{p}_k \leftarrow$ the largest value of $\boldsymbol{p}_i[k]$ for $\boldsymbol{x}_i \in \mathcal{D}_k$
    $\mathcal{R}_k \leftarrow \{\boldsymbol{x}_i \mid (\boldsymbol{x}_i, y_i) \in \mathcal{D}_k, \boldsymbol{p}_i[k] > \tau \times \tilde{p}_k\}$
    $\boldsymbol{\mu}_k \leftarrow$ prototype calculated by Eq. (3)
    $d_i \leftarrow ||\boldsymbol{z}_i - \boldsymbol{\mu}_k||_2$ for $\boldsymbol{x}_i \in \mathcal{D}_k$
    $w_i \leftarrow$ clean prob from GMM fitted on $\{d_i\}_{i=1}^{|\mathcal{D}_k|}$
    $\mathcal{S}_l \leftarrow \mathcal{S}_l \cup \{(\boldsymbol{x}_i, y_i) \in \mathcal{D}_k \mid w_i > 0.5\}$
    $\mathcal{S}_u \leftarrow \mathcal{S}_u \cup \{(\boldsymbol{x}_i, y_i) \in \mathcal{D}_k \mid w_i \leq 0.5\}$
  else
    $\mathcal{S}_l \leftarrow \mathcal{S}_l \cup \{(\boldsymbol{x}_i, y_i) \in \mathcal{D}_k\}$
  end if
end for

Algorithm B4 Clean-Dominant Class Identification

Input: dataset $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, ready-to-select class set $\mathcal{C}$, labeled sample set $\mathcal{S}_l$, and model f
Output: sample set for negative learning $\mathcal{S}_n$

$\mathcal{S}_n \leftarrow \emptyset$
for class $k \in \mathcal{C}$ **do**
  fetch class-wise training dataset $\mathcal{D}_k$
  $\{\hat{y}_i\}_{i=1}^{|\mathcal{D}_k|} \leftarrow f(\boldsymbol{x}_i)$ for $\boldsymbol{x}_i \in \mathcal{D}_k$
  $\hat{y}_k^{mode} \leftarrow$ the majority among $\{\hat{y}_i\}_{i=1}^{|\mathcal{D}_k|}$
  if $\hat{y}_k^{mode} = k$ **then**
    $\mathcal{S}_n \leftarrow \mathcal{S}_n \cup \{(\boldsymbol{x}_i, y_i) \in \mathcal{S}_l \mid y_i = k\}$
  end if
end for

C Proofs

Here we provide the rigorous statements, assumptions, and proofs of Theorems 3.1, 3.2, and 3.4 and Lemma 3.3 in the main text.

Proof of Theorem 3.1. Let the observed margin for sample i be $Z_i = p_i[k] - \alpha \max_{j \neq k} p_i[j]$. We define the binary indicator $\mathcal{E}_i = \mathbb{I}(Z_i \geq 0)$.

Let the empirical average margin be $\Delta_k = \frac{1}{|\mathcal{M}_k|} \sum_{i \in \mathcal{M}_k} Z_i$ and the empirical fraction of non-negative predictions be $r_k = \frac{1}{|\mathcal{M}_k|} \sum_{i \in \mathcal{M}_k} \mathcal{E}_i$.

We first establish a correspondence between Δ_k and r_k. To render the analysis tractable, we consider an idealized margin distribution where negative margin samples are bounded at $-\alpha$ and positive margin samples saturate at 1. Under this simplification, the empirical average is expressed as:

$$\Delta_k = r_k(1) + (1 - r_k)(-\alpha). \tag{15}$$

Rearranging this linear relationship yields the equality:

$$r_k = \frac{\Delta_k + \alpha}{1 + \alpha}. \tag{16}$$

Next, we connect r_k to the true purity. Let $s_k = \mathbb{E}[\mathcal{E}_i \mid i \in \mathcal{M}_k]$. Defining $p_{c,k} = \mathbb{P}(\mathcal{E}_i \mid y_i^* = k, i \in \mathcal{M}_k)$ and $p_{n,k} = \mathbb{P}(\mathcal{E}_i \mid y_i^* \neq k, i \in \mathcal{M}_k)$, the law of total probability gives:

$$s_k = \text{Purity}(\mathcal{M}_k)p_{c,k} + (1 - \text{Purity}(\mathcal{M}_k))p_{n,k}. \tag{17}$$

Using the definition $\omega_k = p_{c,k} - p_{n,k}$, this simplifies to $s_k = p_{n,k} + \omega_k \cdot \text{Purity}(\mathcal{M}_k)$.

Since $\mathcal{E}_i$ are independent Bernoulli variables, by Hoeffding's inequality, with probability at least $1 - \delta$, we have $r_k \leq s_k + \varepsilon_k(\delta)$. Substituting s_k and isolating $\text{Purity}(\mathcal{M}_k)$ provides:

$$\text{Purity}(\mathcal{M}_k) \geq \omega_k^{-1} \left(r_k - p_{n,k} - \varepsilon_k(\delta) \right). \tag{18}$$

Finally, substituting r_k from Eq. (16) and applying the assumption $p_{n,k} \leq 0.5$ yields:

$$\text{Purity}(\mathcal{M}_k) \geq \omega_k^{-1} \left(\frac{\Delta_k + \alpha}{1 + \alpha} - \frac{1}{2} - \varepsilon_k(\delta) \right). \tag{19}$$

□

Proof of Theorem 3.2. Let $\mathcal{D}_k$ denote the set of observed samples for class k with cardinality $n_k = |\mathcal{D}_k|$. Let $\mathcal{D}_k^*$ denote the subset of true clean samples, such that $|\mathcal{D}_k^*| = q_k n_k$, where q_k is the clean proportion. Consequently, the number of noisy samples is $|\mathcal{D}_k \setminus \mathcal{D}_k^*| = (1 - q_k)n_k$.

We define the selection indicator variable $S_i \in \{0, 1\}$ for a sample $\boldsymbol{x}_i$, where $S_i = 1$ if $w_i > 0.5$ (i.e., $\boldsymbol{x}_i \in C_k$) and $S_i = 0$ otherwise. Based on the assumption that clean and noisy samples are well-separated with a selection error bound $\gamma < 1/2$, we denote the true positive rate (selection rate of clean samples) as α_{TP} and the false positive rate (selection rate of noisy samples) as α_{FP}. These rates satisfy the following bounds:

$$\alpha_{\text{TP}} = \mathbb{P}(S_i = 1 | \boldsymbol{x}_i \in \mathcal{D}_k^*) \geq 1 - \gamma, \tag{20}$$

$$\alpha_{\text{FP}} = \mathbb{P}(S_i = 1 | \boldsymbol{x}_i \notin \mathcal{D}_k^*) \leq \gamma. \tag{21}$$

Since $\gamma < 1/2$, it strictly follows that $\alpha_{\text{TP}} \geq 1 - \gamma > 0.5 > \gamma \geq \alpha_{\text{FP}}$, implying $\alpha_{\text{TP}} > \alpha_{\text{FP}}$.

Now, we analyze the purity of the selected set C_k, defined as the ratio of selected clean samples to the total size of C_k. The expected purity is given by:

$$\text{Purity}(C_k) = \frac{\alpha_{\text{TP}} q_k n_k}{\alpha_{\text{TP}} q_k n_k + \alpha_{\text{FP}}(1 - q_k)n_k} = \frac{\alpha_{\text{TP}} q_k}{\alpha_{\text{TP}} q_k + \alpha_{\text{FP}}(1 - q_k)}. \tag{22}$$

To prove that $\text{Purity}(C_k) > q_k$, we examine the inequality:

$$\frac{\alpha_{\text{TP}} q_k}{\alpha_{\text{TP}} q_k + \alpha_{\text{FP}}(1 - q_k)} > q_k. \tag{23}$$

Since $q_k > 0$, we can divide both sides by q_k, reducing the inequality to:

$$\frac{\alpha_{\text{TP}}}{\alpha_{\text{TP}} q_k + \alpha_{\text{FP}}(1 - q_k)} > 1 \implies \alpha_{\text{TP}} > \alpha_{\text{TP}} q_k + \alpha_{\text{FP}}(1 - q_k). \tag{24}$$

Rearranging terms yields $\alpha_{\text{TP}}(1 - q_k) > \alpha_{\text{FP}}(1 - q_k)$. Since $q_k < 1$ (assuming the dataset contains noise), we divide by $(1 - q_k)$ to obtain $\alpha_{\text{TP}} > \alpha_{\text{FP}}$. As established earlier, the error bound assumption $\gamma < 1/2$ guarantees $\alpha_{\text{TP}} > \alpha_{\text{FP}}$. Thus, the inequality holds, proving $\text{Purity}(C_k) > q_k$.

Asymptotically, as $|\mathcal{D}_k| \to \infty$, the Law of Large Numbers ensures the prototype $\boldsymbol{\mu}_k$ converges to the true mean of the clean distribution (given $q_k > 1/2$). The assumption that samples are well separated implies that the overlap between the clean and noisy distributions in the feature space is negligible. In this limit, the Gaussian Mixture Model can perfectly distinguish the modes, implying the selection error $\gamma \to 0$. Consequently, $\alpha_{\text{TP}} \to 1$ and $\alpha_{\text{FP}} \to 0$. Substituting these limits into the derived expressions:

$$\lim_{|\mathcal{D}_k| \to \infty} \frac{|C_k|}{|\mathcal{D}_k|} \geq (1 - 0)q_k = q_k, \tag{25}$$

and

$$\lim_{|\mathcal{D}_k| \to \infty} \text{Purity}(C_k) = \frac{1 \cdot q_k}{1 \cdot q_k + 0 \cdot (1 - q_k)} = 1. \tag{26}$$

□

The following theorem demonstrates that the average *purity* of class labels is critical in the generalization bound. This motivates us to identify clean samples.

Proof of Lemma 3.3. We decompose the excess risk as follows:

$$\begin{aligned} R(\hat{f}) - R(f^*) = &\left[R(\hat{f}) - \hat{R}_{S_c}(\hat{f})\right] \\ &+ \left[\hat{R}_{S_c}(\hat{f}) - \hat{R}_{S_c}(f^*)\right] \\ &+ \left[\hat{R}_{S_c}(f^*) - R(f^*)\right]. \end{aligned} \tag{27}$$

To bound the middle term, note that since $\hat{f}$ minimizes the empirical risk over S, we have:

$$\begin{aligned} &\frac{1}{N}\left(N_c \hat{R}_{S_c}(\hat{f}) + N_n \hat{R}_{S_n}(\hat{f})\right) \\ &\leq \frac{1}{N}\left(N_c \hat{R}_{S_c}(f^*) + N_n \hat{R}_{S_n}(f^*)\right). \end{aligned} \tag{28}$$

Multiplying both sides by N and rearranging to isolate the clean risk yields:

$$N_c\left(\hat{R}_{S_c}(\hat{f}) - \hat{R}_{S_c}(f^*)\right) \leq N_n\left(\hat{R}_{S_n}(f^*) - \hat{R}_{S_n}(\hat{f})\right). \tag{29}$$

By boundedness, $\hat{R}_{S_n}(f^*) \leq L$ and $\hat{R}_{S_n}(\hat{f}) \geq 0$. Thus, the RHS is bounded by $N_n L$, leading to:

$$\hat{R}_{S_c}(\hat{f}) - \hat{R}_{S_c}(f^*) \leq \frac{N_n}{N_c} L. \tag{30}$$

To bound the generalization gaps (the first and third terms in Eq. (27)), we apply a standard Rademacher complexity bound. With probability at least $1-\delta$, for all $f \in \mathcal{F}$:

$$|R(f) - \hat{R}_{S_c}(f)| \leq 2L\mathfrak{R}_{N_c}(\mathcal{F}) + L\sqrt{\frac{\log(2/\delta)}{2N_c}}. \tag{31}$$

Applying this to both $\hat{f}$ and f^*, we obtain:

$$R(\hat{f}) - \hat{R}_{S_c}(\hat{f}) \leq 2L\mathfrak{R}_{N_c}(\mathcal{F}) + L\sqrt{\frac{\log(2/\delta)}{2N_c}}, \tag{32}$$

$$\hat{R}_{S_c}(f^*) - R(f^*) \leq 2L\mathfrak{R}_{N_c}(\mathcal{F}) + L\sqrt{\frac{\log(2/\delta)}{2N_c}}. \tag{33}$$

Substituting these three bounds into Eq. (27), we obtain:

$$R(\hat{f}) - R(f^*) \leq L\frac{N_n}{N_c} + 4L\mathfrak{R}_{N_c}(\mathcal{F}) + 2L\sqrt{\frac{\log(2/\delta)}{2N_c}}. \tag{34}$$

Rearranging yields the claimed result. □

Proof of Theorem 3.4. We begin by defining the generalization bound function $B(N_c, N_n, \delta)$ derived from Lemma 3.3. By substituting the explicit terms, the bound is expressed as:

$$\begin{aligned} B(N_c, N_n, \delta) = &R(f^*) + L\frac{N_n}{N_c} \\ &+ 4L\mathfrak{R}_{N_c}(\mathcal{F}) + 2L\sqrt{\frac{\log(2/\delta)}{2N_c}}. \end{aligned} \tag{35}$$

We compare the value of this bound for the original training set S against the selected subset S_{sel}, characterized by the sample counts N_c^{sel} and N_n^{sel}.

First, we establish the inequality $B(N_c^{\text{sel}}, N_n^{\text{sel}}, \delta) < B(N_c, N_n, \delta)$ for sufficiently large N_c. The strict noise reduction condition implies that the noise-to-clean ratio satisfies $\frac{N_n^{\text{sel}}}{N_c^{\text{sel}}} \leq \frac{N_n}{N_c} - \varepsilon$. Consequently, the contribution of the noise term to the bound decreases by at least $2L\varepsilon$. Conversely, the Rademacher complexity and generalization error terms depend inversely on the clean sample count. Since the selection strategy retains a fraction of clean data such that $N_c^{\text{sel}} \geq cN_c$ for some $c \in (0, 1]$, the reduction in sample size may strictly increase these terms compared to the full dataset. However, by the assumption that $\mathfrak{R}_m(\mathcal{F}) \to 0$ as $m \to \infty$, both the complexity term (which scales as $4L\mathfrak{R}_{N_c}(\mathcal{F})$) and the generalization error term (which scales as $\mathcal{O}(N_c^{-1/2})$) vanish asymptotically. Therefore, the difference in the non-noise terms between the selected and original sets also converges to zero as $N_c \to \infty$. Since the reduction in the noise term is bounded away from zero by the constant $2L\varepsilon$, there exists a threshold N_0 such that for all $N_c > N_0$, the reduction in noise dominates the marginal increase in complexity, yielding a strictly tighter bound for the selected subset.

Moreover, we examine the asymptotic convergence under a consistent selection strategy where $N_c^{\text{sel}}/N_c \to 1$ and $N_n^{\text{sel}}/N_c^{\text{sel}} \to 0$. In the limit as $N_c \to \infty$, the noise term $2L\frac{N_n^{\text{sel}}}{N_c^{\text{sel}}}$ for the selected subset vanishes entirely. Simultaneously, since N_c^{sel} grows at the same rate as N_c, the complexity and generalization error terms for the selected subset converge to those of a dataset containing N_c clean samples. The oracle bound $B(N_c, 0, \delta)$ represents the theoretical risk bound for a model trained on N_c clean samples with zero noise. Because the noise term in S_{sel} vanishes and the clean sample count asymptotically matches S_c, the difference $|B(N_c^{\text{sel}}, N_n^{\text{sel}}, \delta) - B(N_c, 0, \delta)|$ approaches zero. Thus, the bound for the selected subset converges to the oracle bound. □

D Empirical Validity of Theoretical Assumptions

Table D1: Empirical validity of the assumptions underlying Theorem 3.2 and sample selection purity at the first-ready epoch. The values for clean-majority and feature separability indicate the average number of classes satisfying each assumption.

Dataset	Label Noise	Clean-Majority (q_k)		Feature Separability (γ)		Initial Purity → Selected Purity	
		Head	Tail	Head	Tail	Head	Tail
CIFAR-10	Sym-40%	5/5	1/5	5/5	4.8/5	72.95% → 94.90%	41.49% → 86.90%
	Sym-60%	3/5	0/5	4.8/5	4.5/5	56.44% → 87.02%	25.24% → 69.12%
CIFAR-100	Sym-40%	50/50	8/50	46.7/50	35.5/50	68.47% → 90.44%	41.05% → 73.82%
	Sym-60%	24/50	0/50	37/50	21.5/50	49.16% → 79.16%	24.23% → 54.28%

We empirically examine how often the assumptions underlying Theorem 3.2 are satisfied. We evaluate each class at its first-ready epoch on CIFAR-10 and CIFAR-100 under 40% and 60% symmetric noise with an imbalance factor of $\rho = 0.1$. Clean-majority requires $q_k > 1/2$, meaning that more than half of the samples with observed label k are clean, whereas feature separability requires $\gamma = \max(1 - \alpha_{\text{TP}}, \alpha_{\text{FP}}) < 1/2$, where α_{TP} and α_{FP} are the selection rates of clean and noisy samples, respectively. The head and tail groups refer to the first and latter halves of the classes, respectively. Table D1 reports the average number of classes satisfying each assumption across repeated runs, together with the purity before and after sample selection.

Clean-majority was satisfied much more frequently for head classes than for tail classes. Under 40% symmetric noise, all head classes satisfied clean-majority, whereas only 1/5 CIFAR-10 tail classes and 8/50 CIFAR-100 tail classes satisfied the assumption. Under 60% symmetric noise, no tail class satisfied clean-majority, and the number of satisfying head classes also decreased. Feature separability was satisfied for most CIFAR-10 classes even under 60% symmetric noise. On CIFAR-100, the number of satisfying classes decreased from 46.7/50 to 37/50 for head classes and from 35.5/50 to 21.5/50 for tail classes as the noise rate increased.

These results show that the theoretical assumptions are more frequently satisfied for head classes than for tail classes, particularly under severe label noise. Nevertheless, the average selected purity increased over the initial purity for every dataset and class group reported in Table D1.

E Detailed Experimental Setup

E.1 Benchmark Datasets

We validated the proposed method on two benchmark datasets: CIFAR-10 and CIFAR-100 [23]. Following previous studies [41, 31, 48], we first simulated class imbalance and then injected synthetic label noise into the datasets.

Specifically, the class imbalance was introduced using an exponential decay function $n_1 \cdot \rho^{(k-1)/(K-1)}$, where $\rho = n_K/n_1$ is an imbalance factor, which indicates the ratio of sample sizes between the least and most frequent classes, and K is the total number of classes.

To corrupt the imbalanced dataset, we injected symmetric (Sym), asymmetric (Asym), or instance-dependent (IDN) label noise according to a predefined noise ratio. In the symmetric setting, each label is randomly flipped to one of all classes with uniform probability. In contrast, asymmetric noise is generated by flipping labels only between certain class pairs, following the configuration in Lu et al. [31]. For the instance-dependent noise, we follow Xia et al. [43] and corrupt labels with probabilities that depend on each sample.

For evaluation, we adopted the protocol used in prior works [41, 31, 6], where models are trained on noisy and class-imbalanced training sets but evaluated on clean and balanced test sets.

E.2 Real-World Datasets

We further evaluated the proposed method on two real-world datasets: Clothing1M [44] and Animal-10N [34].

Clothing1M consists of one million clothing images across 14 classes, which include noisy labels that occurred during automatic data collection from online shopping websites. Following the protocol in Kim et al. [21], we first constructed a balanced subset of 120k images by random sampling, and then simulated class imbalance in the same manner as in the benchmark datasets. For evaluation, we used the manually verified test set of 10k clean samples provided by Xiao et al. [44].

Animal-10N consists of 50k training images from 10 animal categories with visually similar appearances, along with 10k clean test images. In this dataset, noisy labels naturally arise due to human annotation errors during the data collection process.

As with the benchmark datasets, we trained the model on the noisy labeled and class-imbalanced training sets and evaluated it on clean and balanced test sets.

E.3 Implementation

Following Lu et al. [31], we trained PreAct ResNet-18 for 100 epochs on the benchmark datasets. For the real-world datasets, we adopted the experimental setups used in Kim et al. [21] and Song et al. [34]. Specifically, we trained ResNet-50 for 80 epochs on the Clothing1M, and VGG-19 for 100 epochs on the Animal-10N.

For sample selection methods, including the proposed method, as well as Co-teaching [16], JoCoR [40], DivideMix [27], RoLT [41], TABASCO [31], and CBS [29], we incorporated a warm-up phase during training. Here, the warm-up period was set to 10 epochs for CIFAR-10, Clothing1M, and Animal-10N, and 30 epochs for CIFAR-100.

We repeated all experiments three times for both our method and the baselines, and reported the average classification accuracy throughout the paper.

For the comparison methods that jointly handle label noise and class imbalance [5, 41, 31, 29, 6], we used optimal hyperparameters reported in their papers. For the other baselines [4, 32, 16, 40, 27], we tuned their own hyperparameters.

All experiments except for ImageNet-LT and iNaturalist 2018 were executed using the PyTorch platform on a system with an Intel Core i9-10900X CPU at 3.70 GHz, 192 GB RAM, and a GeForce RTX 4090 24GB GPU. The experiments on ImageNet-LT and iNaturalist 2018 were conducted with an Intel Xeon Platinum 8462Y+ CPU at 4.10 GHz, 2 TB RAM, and an H200 141GB GPU.

F Comparison with Additional Baselines

We compared RASS with the additional baselines H2E [46], SFA [26], and SSBL [47] on CIFAR-10 under supervised and semi-supervised settings. RASS achieved the best performance in all eight supervised settings and all four asymmetric noise settings under the semi-supervised objective, while SFA performed best in the four symmetric noise settings under the semi-supervised objective. Across both training objectives, RASS ranked first in all asymmetric noise settings, where label corruption is concentrated between specific class pairs. These results demonstrate that readiness-aware selection is particularly beneficial when class discrimination is more difficult.

Table F1: Test accuracies on CIFAR-10 with various label noises and imbalance factors under the supervised and semi-supervised objectives. The best and second-best results are highlighted in **bold** and underlined, respectively.

Datasets	Obj.	ρ	0.1		0.05		0.1		0.05	
		Label Noise	Sym-40%	Sym-60%	Sym-40%	Sym-60%	Asym-20%	Asym-40%	Asym-20%	Asym-40%
CIFAR-10	Sup.	H2E	73.65±0.72	57.67±1.58	65.86±1.52	48.68±1.10	80.30±0.55	65.25±2.87	76.54±1.20	65.10±1.01
		SFA	80.68±0.85	70.27±0.07	74.90±0.53	61.17±0.61	81.94±0.56	68.75±0.80	72.35±3.34	66.58±0.36
		SSBL	78.20±0.07	65.12±1.20	72.29±0.71	57.25±1.46	83.38±0.73	69.33±3.12	77.39±0.44	66.75±2.85
		RASS	**83.35±0.44**	**73.64±0.31**	**78.96±0.26**	**66.64±0.39**	**86.36±0.05**	**81.23±0.40**	**83.73±0.34**	**77.53±0.56**
	Semi.	H2E	74.83±1.12	57.57±4.61	67.06±2.28	47.60±1.16	79.64±0.82	65.88±4.36	75.07±0.35	65.81±0.88
		SFA	**86.63±0.22**	**84.99±0.52**	**82.10±1.32**	**80.69±0.47**	82.22±0.67	72.69±0.45	74.45±2.57	69.28±0.39
		SSBL	81.43±0.42	77.34±0.90	75.18±0.53	69.80±0.63	82.56±0.40	71.14±3.10	76.68±2.03	65.34±4.07
		RASS	84.50±0.16	83.89±0.08	81.87±0.10	78.18±0.66	**83.28±0.62**	**85.53±0.33**	**78.27±0.05**	**76.24±0.64**

G Evaluation on Large-Scale Datasets

To further evaluate the scalability and practical applicability of RASS beyond the CIFAR benchmarks, we conduct additional experiments on Tiny-ImageNet, ImageNet-LT, and iNaturalist 2018. These datasets contain substantially more classes than CIFAR-10/100 and allow us to evaluate RASS in large-scale settings with different forms of class imbalance and label noise.

Tiny-ImageNet Tiny-ImageNet contains 200 classes and provides a larger-scale benchmark than CIFAR-10 and CIFAR-100 datasets. We evaluate RASS with an imbalance factor of $\rho = 0.1$ under 40% symmetric noise and 20% asymmetric noise. All methods use the same ViT architecture with nine transformer blocks and 12 attention heads and are trained from scratch. As shown in Table G1, RASS achieved the best Top-1 and Top-5 accuracies under asymmetric noise for both the supervised and semi-supervised objectives. Under symmetric noise, RoLT achieved the best performance, while RASS remained competitive.

ImageNet-LT ImageNet-LT contains 1,000 classes with a long-tailed class distribution. We evaluate RASS under symmetric noise at rates of 40% and 60%, asymmetric noise at rates of 20% and 40%, and instance-dependent noise (IDN) at rates of 20% and 40%. We compare RASS with RoLT and TABASCO, the two strongest sample selection baselines in our main experiments. All methods use the same CLIP-pretrained ViT image encoder with 12 transformer blocks and 12 attention heads, and all model parameters are trained end-to-end for 50 epochs. As shown in Table G2, RASS achieved the best performance in most settings under both training objectives. In particular, RASS achieved the best Top-1 and Top-5 accuracies under Sym-60%, Asym-20%, Asym-40%, and IDN-40% for both objectives, demonstrating strong performance especially under higher levels of label noise.

iNaturalist 2018 iNaturalist 2018 contains 8,142 fine-grained classes with a naturally long-tailed class distribution, providing a more realistic class-imbalanced setting. We evaluate RASS under IDN-20% and IDN-40%, generated using the same protocol as in Appendix E.1. All methods use ResNet-50 as the backbone and are trained from scratch for 100 epochs on the original class distribution without artificially introducing class imbalance. As shown in Table G3, RoLT achieved the best performance under IDN-20%, while RASS achieved the best Top-1 and Top-5 accuracies under IDN-40% for both training objectives. These results show that RASS remains effective under natural class imbalance and becomes particularly beneficial as label noise becomes more severe.

Table G1: Top-1 and Top-5 test accuracies on Tiny-ImageNet with an imbalance factor of $\rho = 0.1$. The best and second-best results are highlighted in **bold** and underlined, respectively.

(a) Supervised objective

Method	Sym-40%		Asym-20%	
	Top-1	Top-5	Top-1	Top-5
HAR	18.35	37.10	25.85	47.10
RoLT	**27.06**	**48.37**	31.30	54.77
TABASCO	26.57	48.23	31.01	54.00
CBS	24.21	45.25	25.41	47.87
RobustLA	25.18	46.52	27.95	48.73
RASS	25.86	46.56	**32.15**	**55.63**

(b) Semi-supervised objective

Method	Sym-40%		Asym-20%	
	Top-1	Top-5	Top-1	Top-5
HAR	–	–	–	–
RoLT	30.05	**53.01**	33.14	55.91
TABASCO	**30.18**	52.71	31.92	53.61
CBS	26.19	47.03	26.16	46.94
RobustLA	–	–	–	–
RASS	28.39	49.68	**34.19**	**58.32**

Table G2: Top-1 and Top-5 test accuracies on ImageNet-LT under symmetric, asymmetric, and instance-dependent noise. The best and second-best results are highlighted in **bold** and underlined, respectively.

(a) Supervised objective

Method	Sym-40%		Sym-60%		Asym-20%		Asym-40%		IDN-20%		IDN-40%	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
RoLT	**39.54**	**65.86**	31.96	51.57	44.97	70.66	32.18	56.82	46.00	73.26	34.37	60.15
TABASCO	37.56	64.54	30.74	50.81	45.91	72.61	32.33	57.29	**46.16**	**74.21**	34.80	61.57
RASS	37.86	64.68	**32.56**	**52.81**	**47.25**	**73.61**	**34.42**	**59.92**	46.06	73.51	**35.35**	**62.07**

(b) Semi-supervised objective

Method	Sym-40%		Sym-60%		Asym-20%		Asym-40%		IDN-20%		IDN-40%	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
RoLT	**42.57**	**70.74**	32.19	53.86	47.88	77.83	36.56	62.83	48.43	77.77	41.19	67.98
TABASCO	41.75	69.15	33.14	55.21	49.14	77.81	36.88	64.39	47.64	76.80	41.28	68.01
RASS	42.36	70.56	**33.83**	**56.89**	**51.06**	**80.26**	**39.47**	**68.64**	**49.48**	**79.50**	**42.52**	**71.59**

Table G3: Top-1 and Top-5 test accuracies on iNaturalist 2018 under instance-dependent noise. The best and second-best results are highlighted in **bold** and underlined, respectively.

(a) Supervised objective

Method	IDN-20%		IDN-40%	
	Top-1	Top-5	Top-1	Top-5
RoLT	**47.25**	**68.32**	34.42	54.85
TABASCO	45.63	67.24	33.80	54.46
RASS	46.66	67.70	**34.63**	**55.08**

(b) Semi-supervised objective

Method	IDN-20%		IDN-40%	
	Top-1	Top-5	Top-1	Top-5
RoLT	**48.96**	**69.44**	37.03	56.71
TABASCO	47.29	68.64	33.79	53.76
RASS	48.58	69.26	**37.39**	**57.13**

H Analysis of Model Design and Hyperparameters

H.1 Additional Sensitivity Analysis

We conducted a sensitivity analysis on four key hyperparameters: the confidence margin α, which determines class readiness; the confidence threshold τ, which is used to construct class prototypes; and the loss weights λ_1 and λ_2, which control the negative loss and the penalty loss, respectively. Figures H1(a) and H1(b) show the results under the supervised and semi-supervised objectives. In both figures, the dashed lines indicate the hyperparameter values used in the comparisons with baseline methods in the main text.

Across both training objectives, the confidence margin α and the penalty loss weight λ_2 performed consistently well when set to 3 and 1, respectively. Since too small α may cause insufficiently learned classes to be mistakenly regarded as ready-to-select, a large enough α is required to perform reliable

sample selection. For λ_2, a large value can lead to overly uniform predictive distributions, while a small value can result in severely biased distributions toward majority classes. We observed that the model consistently achieved effective performance when λ_2 is set to 1.

The other two hyperparameters, τ and λ_1, exhibited objective-dependent sensitivity. For the confidence threshold τ, model performance was relatively stable under the supervised objective. However, under the semi-supervised objective, particularly in challenging scenarios such as Sym-60% and Asym-40% with an imbalance factor of 0.05, increasing τ led to a notable drop in classification performance. A high τ encourages the use of only highly confident samples to construct prototypes, improving their purity but reducing their coverage. While a threshold of $\tau = 0.8$ still yielded decent results, our findings highlight that broader prototype coverage is important under severe corruption and class imbalance in the semi-supervised objective.

The negative loss weight λ_1 also showed objective-dependent sensitivity. For the supervised objective, smaller values (e.g., 1e-3) generally showed better performance. For the semi-supervised objective, although the clear trend was not observed, performance often improved when λ_1 was set to values larger than 1e-3.

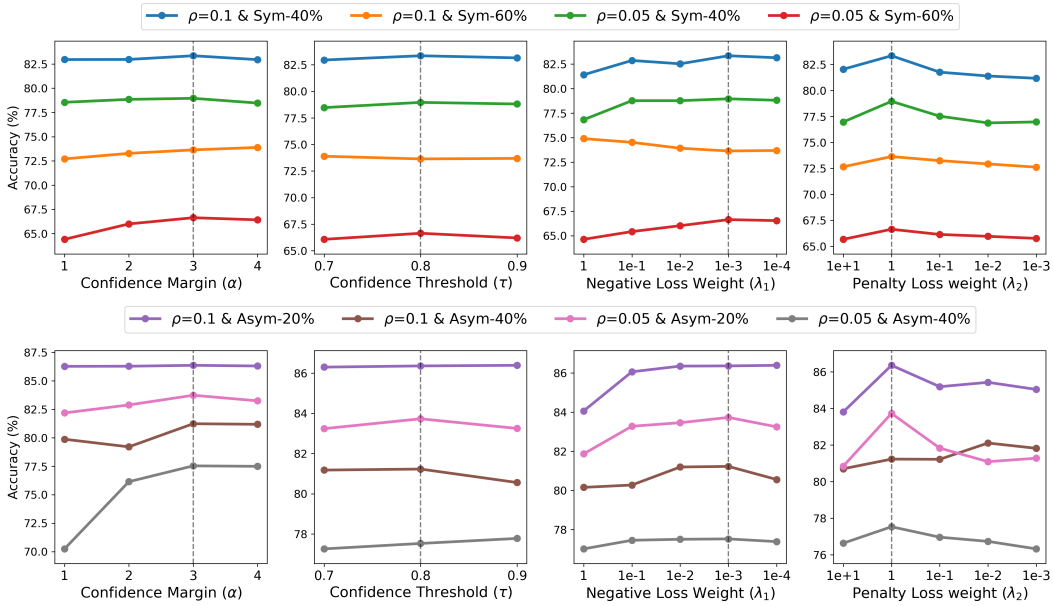
To examine the sensitivity of RASS on a dataset with more classes, we evaluated $\alpha \in \{1, 2, 3, 4\}$ and $\tau \in \{0.7, 0.8, 0.9\}$ on CIFAR-100 in Figure H2. Among these settings, $\alpha = 3$ achieved the best performance in nine cases, while $\alpha = 2$ was best in the remaining seven. In particular, $\alpha = 3$ was preferred in most asymmetric-noise settings, indicating that more challenging label-noise conditions may benefit from a more conservative readiness criterion. Nevertheless, the performance difference between $\alpha = 2$ and $\alpha = 3$ was not substantial. Similarly, although the best value of τ varied across settings, the performance remained relatively stable within the evaluated range.

H.2 Additional Ablation Study

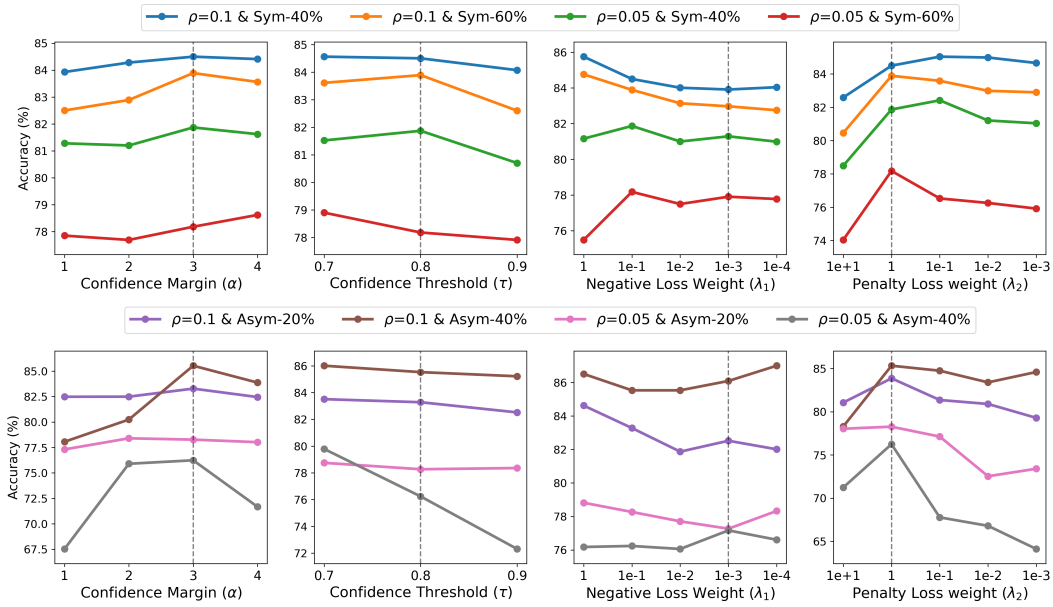
Negative Learning We conducted additional ablation studies on CIFAR-10 under Sym-60% noise with imbalance factors of $\rho = 0.1$ and $\rho = 0.05$. We incorporated negative learning into RASS to reduce confusion of the model and facilitate class readiness. As shown in Figure H3, RASS increases the ready-class ratio faster than its variant without negative learning (w/o NL). Such a tendency is more clearly observed under the severe imbalance factor $\rho = 0.05$. These results suggest that the proposed negative learning helps unready classes become ready for selection under the imbalanced datasets. As a result, our negative learning scheme contributes to improving the quality of sample selection, as shown in Figure 3, and enhances the generalization performance of the model, as reported in Table 4.

Handling Unready Classes Table H1 presents the results of different strategies for handling unready-class samples. In RASS, samples from unready classes are included in the labeled set $\mathcal{S}_l$. Treating these samples as unlabeled samples $\mathcal{S}_u$ can be harmful depending on the training objective. Under the supervised objective, samples in $\mathcal{S}_u$ are not used for training, which removes the opportunity for the model to learn discriminative features for unready classes. Under the semi-supervised objective, samples in $\mathcal{S}_u$ are trained with pseudo-labels, which can be unreliable for unready classes whose features are not sufficiently learned. As shown in Table H1, treating unready-class samples as $\mathcal{S}_u$ led to substantial performance degradation across all settings. For example, under the supervised objective with Sym-60% noise, the accuracy decreased from 73.64% to 45.85% when $\rho = 0.1$, and from 66.64% to 41.32% when $\rho = 0.05$. Similar degradation was observed under the semi-supervised objective, where the accuracy decreased from 83.89% to 45.85% and from 78.18% to 41.34%, respectively.

We further evaluated down-weighting the supervised loss for unready-class samples while retaining them in $\mathcal{S}_l$. Although moderate down-weighting with a weight of 0.5 resulted in smaller performance degradation than treating the samples as $\mathcal{S}_u$, it consistently underperformed RASS across all noise and imbalance settings. A stronger down-weighting with a weight of 0.1 further degraded performance. These results indicate that retaining sufficient supervision for unready classes is important for learning discriminative features and reaching readiness, supporting our design choice of keeping unready-class samples in $\mathcal{S}_l$ without down-weighting.

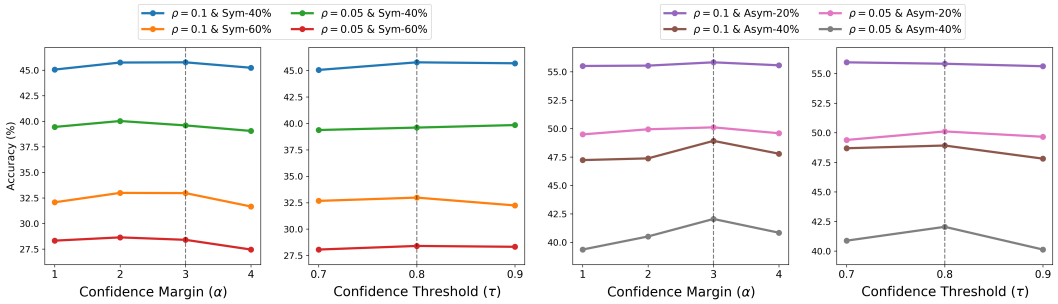


(a) Supervised objective

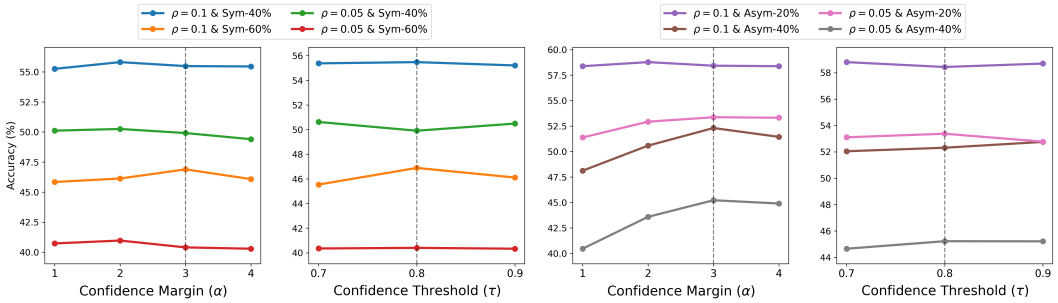


(b) Semi-supervised objective

Figure H1: Sensitivity analysis for the confidence margin α, confidence threshold τ, the weights for the negative loss λ_1, and penalty loss λ_2 on the CIFAR-10 dataset under the (a) supervised and (b) semi-supervised objectives. In each figure, the first and second rows represent the results under the symmetric and asymmetric noises, respectively, and the dashed lines indicate the values used to compare RASS with other baselines.

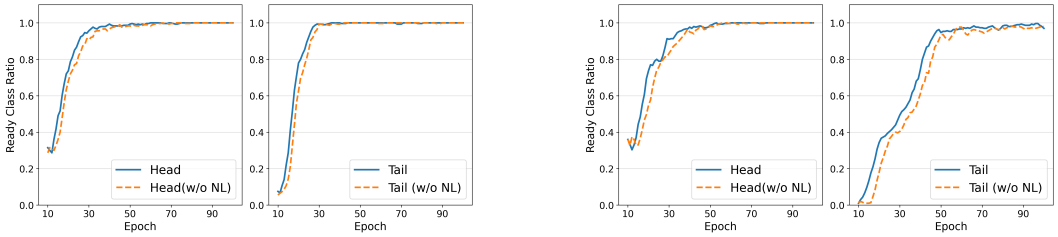


(a) Supervised objective



(b) Semi-supervised objective

Figure H2: Sensitivity analysis for the confidence margin α and confidence threshold τ on the CIFAR-100 dataset under the (a) supervised and (b) semi-supervised objectives. In each figure, the first two and last two columns represent the results under the symmetric and asymmetric noises, respectively, and the dashed lines indicate the values used to compare RASS with other baselines.



(a) Imbalance Factor: 0.1

(b) Imbalance Factor: 0.05

Figure H3: Readiness dynamics of head and tail classes on CIFAR-10 under Sym-60% noise. *Ready Class Ratio* denotes the proportion of classes identified as ready-to-select. Solid lines indicate RASS, and dashed lines indicate the variant without negative learning (w/o NL).

Table H1: Ablation study of strategies for handling unready-class samples. We compare RASS with variants that treat unready-class samples as unlabeled or down-weight their supervised loss.

Obj.	Unready-Class Samples	Sym-60%		Asym-40%	
		$\rho = 0.1$	$\rho = 0.05$	$\rho = 0.1$	$\rho = 0.05$
Sup	Treated as $\mathcal{S}_l$ (RASS)	**73.64** ± 0.31	**66.64** ± 0.39	**81.23** ± 0.40	**77.53** ± 0.56
	Treated as $\mathcal{S}_u$	45.85 ± 1.81	41.32 ± 2.24	59.16 ± 4.00	54.07 ± 5.26
	Down-weighted in $\mathcal{S}_l$ (0.5)	72.80 ± 0.35	65.81 ± 0.90	80.48 ± 0.46	76.11 ± 0.73
	Down-weighted in $\mathcal{S}_l$ (0.1)	66.68 ± 0.22	56.36 ± 1.31	75.35 ± 3.36	68.13 ± 1.13
Semi	Treated as $\mathcal{S}_l$ (RASS)	**83.89** ± 0.08	**78.18** ± 0.66	**85.53** ± 0.33	**76.24** ± 0.64
	Treated as $\mathcal{S}_u$	45.85 ± 1.81	41.34 ± 2.22	60.53 ± 1.88	57.62 ± 2.50
	Down-weighted in $\mathcal{S}_l$ (0.5)	82.06 ± 0.43	71.57 ± 2.85	77.78 ± 4.43	69.04 ± 6.86
	Down-weighted in $\mathcal{S}_l$ (0.1)	47.83 ± 1.29	43.82 ± 0.75	68.95 ± 0.56	61.42 ± 1.70

I Results under Instance-Dependent Noise

Here, we present the results under instance-dependent noise. As shown in Tables I1 and I2, RASS achieved the best average rank across both datasets and training objectives. Although TABASCO showed strong performance in relatively mild settings under IDN-20% noise with $\rho = 0.1$, RASS achieved the best performance in most of the remaining cases. The superiority of RASS was particularly clear when the dataset was severely imbalanced and heavily corrupted. For example, under IDN-40% noise with $\rho = 0.05$, RASS achieved the best performance on both CIFAR-10 and CIFAR-100 under the supervised and semi-supervised objectives. In the supervised and semi-supervised objectives, RASS improved the second-best accuracy on CIFAR-10 by 6.13% and 2.35%, respectively. These results indicate that RASS remains effective even when label corruption depends on individual instances, demonstrating that the proposed readiness-aware sample selection strategy is not limited to symmetric or asymmetric label noise.

Table I1: Test accuracies on CIFAR-10 and CIFAR-100 datasets with instance-dependent noise and imbalance factors under the supervised objective. The best and second-best results are highlighted in **bold** and underlined, respectively.

Datasets	ρ	0.1		0.05		Avg
	Label Noise	IDN-20%	IDN-40%	IDN-20%	IDN-40%	Rank
CIFAR-10	CE	81.55 ± 1.00	72.75 ± 0.87	77.70 ± 0.79	64.81 ± 1.55	6.75
	LDAM	82.44 ± 0.31	71.65 ± 0.59	78.23 ± 0.44	63.05 ± 1.08	7.25
	LA	81.52 ± 0.61	70.61 ± 0.17	76.79 ± 0.81	63.42 ± 1.40	8.50
	Co-teaching	80.91 ± 0.23	75.37 ± 0.63	75.07 ± 0.55	56.15 ± 4.82	9.25
	JoCoR	79.76 ± 0.64	69.90 ± 1.24	75.57 ± 0.38	56.58 ± 2.17	10.50
	DivideMix	84.98 ± 0.08	76.28 ± 0.96	81.70 ± 0.39	70.57 ± 0.32	3.75
	HAR	79.62 ± 0.64	72.35 ± 1.23	76.34 ± 1.53	65.74 ± 2.75	8.00
	RoLT	83.89 ± 0.38	72.26 ± 0.28	77.79 ± 1.18	65.54 ± 2.85	6.25
	TABASCO	**87.25** ± 0.32	81.42 ± 0.26	82.89 ± 0.37	74.20 ± 0.18	1.75
	CBS	68.12 ± 1.86	67.20 ± 0.45	62.20 ± 2.83	57.72 ± 1.49	11.50
	RobustLA	85.52 ± 0.23	80.46 ± 1.44	82.01 ± 0.10	68.50 ± 0.67	3.25
	RASS	87.12 ± 0.12	**83.36** ± 0.56	**84.46** ± 0.23	**80.33** ± 0.72	**1.25**
CIFAR-100	CE	48.31 ± 0.13	37.31 ± 0.44	43.65 ± 0.19	33.39 ± 0.28	9.75
	LDAM	49.58 ± 0.18	39.11 ± 0.76	44.81 ± 0.69	34.52 ± 0.47	8.00
	LA	48.76 ± 0.56	37.26 ± 0.27	44.10 ± 0.45	33.31 ± 0.34	9.75
	Co-teaching	52.13 ± 0.17	43.77 ± 0.30	45.30 ± 0.27	37.78 ± 0.51	5.25
	JoCoR	52.34 ± 0.76	44.01 ± 0.25	47.00 ± 0.95	39.36 ± 0.25	3.75
	DivideMix	52.36 ± 0.20	40.14 ± 0.37	46.68 ± 0.37	36.08 ± 0.45	5.50
	HAR	42.23 ± 1.20	32.47 ± 0.54	38.20 ± 0.66	28.13 ± 0.52	12.00
	RoLT	53.72 ± 0.10	41.71 ± 0.12	47.42 ± 0.25	37.63 ± 0.34	4.00
	TABASCO	**55.33** ± 0.40	45.70 ± 0.28	**49.51** ± 0.09	40.18 ± 0.80	**1.50**
	CBS	43.75 ± 0.44	38.25 ± 0.48	38.55 ± 0.57	32.96 ± 0.61	10.50
	RobustLA	51.32 ± 0.81	40.19 ± 1.10	45.80 ± 0.73	35.82 ± 0.53	6.50
	RASS	54.64 ± 0.47	**46.31** ± 0.63	48.60 ± 0.19	**40.55** ± 0.95	**1.50**

Table I2: Test accuracies on CIFAR-10 and CIFAR-100 datasets with instance-dependent noise and imbalance factors under the semi-supervised objective. The best and second-best results are highlighted in **bold** and underlined, respectively.

Datasets	ρ	0.1		0.05		Avg
	Label Noise	IDN-20%	IDN-40%	IDN-20%	IDN-40%	Rank
CIFAR-10	Co-teaching	77.37 ± 0.41	76.77 ± 0.28	72.92 ± 0.07	65.88 ± 1.45	4.75
	JoCoR	80.10 ± 1.53	69.62 ± 13.20	77.03 ± 3.09	51.96 ± 1.79	5.25
	DivideMix	82.55 ± 0.27	84.11 ± 0.06	77.22 ± 0.76	78.48 ± 4.56	3.00
	RoLT	77.99 ± 0.55	68.20 ± 0.52	71.19 ± 4.06	57.83 ± 1.12	6.00
	TABASCO	**84.25** ± 0.16	86.08 ± 0.25	79.07 ± 1.81	81.16 ± 0.57	1.75
	CBS	72.03 ± 0.79	71.09 ± 0.57	69.46 ± 0.44	61.85 ± 1.36	6.00
	RASS	84.04 ± 0.68	**86.66** ± 0.21	**80.56** ± 0.18	**83.51** ± 0.11	**1.25**
CIFAR-100	Co-teaching	54.40 ± 0.51	44.82 ± 1.13	48.56 ± 0.40	38.94 ± 0.93	5.25
	JoCoR	54.85 ± 0.17	42.82 ± 0.44	46.57 ± 0.57	37.89 ± 0.58	6.00
	DivideMix	58.32 ± 0.21	49.19 ± 0.07	51.62 ± 0.34	43.31 ± 0.32	4.00
	RoLT	59.79 ± 0.15	50.79 ± 0.46	53.04 ± 0.21	44.57 ± 0.40	2.75
	TABASCO	**60.17** ± 0.49	52.97 ± 0.34	53.93 ± 0.06	47.42 ± 0.72	1.75
	CBS	43.75 ± 0.44	42.93 ± 0.06	38.55 ± 0.57	36.21 ± 0.20	6.75
	RASS	59.14 ± 0.23	**53.32** ± 0.23	**54.05** ± 0.13	**47.69** ± 0.43	**1.50**

J Analysis of Head and Tail Classes

J.1 Class Readiness

Since model predictions can be biased toward head classes in class-imbalanced datasets, we analyze how such prediction bias affects class readiness estimation, particularly for tail classes.

Delayed Readiness Class imbalance can bias the predicted probability distribution toward head classes, thereby reducing the probability margins for tail classes. As a result, readiness can be delayed for tail classes. Table J1 reports the average first-ready epoch of head and tail classes on CIFAR-10. When the imbalance factor decreased from 0.1 to 0.05, the average first-ready epoch of tail classes increased from 10.6 to 13.1 under 40% symmetric noise and from 17.0 to 28.8 under 60% symmetric noise. In contrast, the first-ready epoch of head classes remained nearly unchanged. These results indicate that severe class imbalance mainly delays readiness for tail classes.

Table J1: Average first-ready epoch of head and tail classes on CIFAR-10 under symmetric label noise and different imbalance factors.

Label Noise	ρ	Head	Tail
Sym-40%	0.1	10.5	10.6
	0.05	10.7	13.1
Sym-60%	0.1	14.9	17.0
	0.05	14.2	28.8

False Readiness Prediction bias can cause a false-ready decision for severely corrupted tail classes. Head-class samples whose observed labels are corrupted to tail class k may be predicted as k with high confidence and dominate $\mathcal{M}_k$. The resulting probability margin can satisfy Eq. (1), even though $\mathcal{M}_k$ does not reliably represent clean samples from class k. We observed no such cases on CIFAR-10. Even on CIFAR-100 with 60% symmetric noise, false-ready decisions affected only 2.1% of classes at an imbalance factor of 0.1 and 5.1% at an imbalance factor of 0.05. All cases were confined to tail classes in which clean samples accounted for only 10% to 20% of the observed class. Moreover, these decisions were temporary. As training continued with classes correctly identified as ready, every false-ready class subsequently transitioned to true readiness.

J.2 Classification Performance

We reported the accuracies for head and tail classes in Table J2. Head and tail classes are defined as the first and last half of the classes sorted by the number of training samples with clean labels.

RASS achieved the best average rank for both head and tail classes, with average ranks of 1.63 and 2.13, respectively. For head classes, RASS obtained the best or second-best performance in all cases. RASS also demonstrated effective performance on tail classes, achieving the best or second-best results in several challenging cases, such as Sym-60% and Asym-40% under severe imbalance. In contrast, sample selection methods that do not consider class imbalance, such as Co-teaching, JoCoR, and DivideMix, tended to show biased performance toward head classes under severe imbalance. For example, under Sym-60% noise with $\rho = 0.05$, Co-teaching, JoCoR, and DivideMix showed relatively high head-class accuracies compared to their tail-class accuracies, indicating that their sample selection results can be dominated by majority classes. Other approaches that are designed to handle both class imbalance and noisy labels, such as TABASCO and RobustLA, achieved strong performance on tail classes, whereas their performance on head classes was relatively limited. These results indicate that the strong performance of RASS is not driven by improvements concentrated in a specific class group. Instead, RASS consistently improves performance across both head and tail classes, demonstrating its effectiveness under noisy labeled datasets with class imbalance.

Table J2: Head and tail class accuracies on CIFAR-10 under different imbalance factors and label noise settings. The best and second-best results are highlighted in **bold** and underlined, respectively.

Method	$\rho = 0.1$				$\rho = 0.05$				$\rho = 0.1$				$\rho = 0.05$				Avg. Rank	
	Sym-40%		Sym-60%		Sym-40%		Sym-60%		Asym-20%		Asym-40%		Asym-20%		Asym-40%			
	Head	Tail	Head	Tail	Head	Tail	Head	Tail	Head	Tail	Head	Tail	Head	Tail	Head	Tail	Head	Tail
CE	77.06	72.67	65.00	63.66	77.22	60.37	68.73	39.08	82.89	81.19	68.57	67.73	84.11	71.45	72.20	58.54	6.63	8.88
LA	76.37	76.35	63.47	67.14	71.45	70.11	60.36	53.48	83.37	82.29	79.27	62.18	79.86	76.04	**79.54**	56.64	7.13	7.50
LDAM	76.02	75.61	64.14	60.90	69.97	62.57	57.55	47.61	85.61	81.05	71.21	67.55	82.97	76.15	67.63	61.47	8.00	7.88
Co-teaching	70.34	77.06	55.05	70.89	68.35	63.80	**74.08**	23.48	76.72	80.85	69.47	62.99	76.42	63.48	72.85	44.71	8.88	9.00
JoCoR	**85.13**	78.31	62.04	25.31	82.61	63.28	70.11	11.59	85.41	83.59	61.85	54.77	86.19	76.09	60.35	51.72	5.63	9.13
DivideMix	81.85	82.11	**72.45**	66.69	78.77	77.09	64.02	53.83	84.23	84.93	73.07	69.93	84.58	78.68	76.57	63.49	4.13	4.75
TABASCO	82.05	83.23	70.09	72.01	78.70	77.38	63.53	59.69	83.41	**87.35**	72.83	82.41	77.99	81.43	67.76	**76.78**	6.50	2.25
RoLT	79.67	83.34	69.57	69.23	74.41	74.44	65.35	55.31	80.73	80.57	76.78	58.53	78.14	75.13	78.45	45.81	6.63	7.38
HAR	75.91	71.89	68.12	59.49	74.67	59.82	62.59	42.21	82.51	80.36	68.87	72.11	81.50	74.24	71.50	62.09	8.38	8.88
Robust LA	80.56	**84.88**	71.39	73.47	75.99	**79.77**	67.39	62.85	85.19	86.98	**81.19**	68.88	82.43	**82.13**	75.31	59.61	4.38	2.50
CBS	63.09	82.27	71.30	69.67	57.63	73.25	67.19	56.45	76.00	61.25	64.35	61.20	59.31	64.22	56.90	56.14	10.13	7.75
RASS	83.20	83.50	72.10	**75.19**	**82.63**	75.29	70.19	**63.10**	**86.41**	86.32	79.83	**82.62**	**86.21**	81.25	79.03	76.03	**1.63**	**2.13**

K Sample Selection Results on CIFAR-100

We provide sample selection results on CIFAR-100 under the supervised objective. We report selection precision (i.e., purity), recall, and F1-score in Table K1. Each metric was computed at every epoch and then averaged across the training epochs. The results demonstrate that RASS achieved the highest selection F1-score in six of the eight settings. Under symmetric noise, methods that do not account for class imbalance, including Co-teaching, JoCoR, and DivideMix, also showed competitive selection performance. Under the more challenging asymmetric noise, methods designed to handle class imbalance, such as RoLT and TABASCO, performed effectively. Among all sample selection approaches, RASS maintained a strong balance between selection precision and recall across different imbalance factors and noise conditions.

Table K1: Sample selection precision (P), recall (R), and F1-score (F1) on CIFAR-100 under the supervised objective. Each metric was computed at every epoch and averaged across the training epochs. The best F1-score in each setting is highlighted in **bold**.

(a) Symmetric noise

Method	$\rho = 0.1$						$\rho = 0.05$					
	Sym-40%			Sym-60%			Sym-40%			Sym-60%		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Co-teaching	80.1	78.7	78.6	68.3	67.4	66.8	79.4	78.3	77.7	57.4	54.4	54.0
JoCoR	83.0	80.9	**81.7**	66.6	67.3	66.4	81.6	80.2	80.3	58.7	59.1	58.2
DivideMix	78.9	84.4	81.1	57.3	83.0	66.4	77.9	86.6	**81.1**	52.1	68.4	57.0
RoLT	74.4	90.0	80.3	56.0	90.9	66.3	70.4	92.2	77.1	47.2	85.3	55.8
TABASCO	78.6	87.5	81.0	59.3	81.3	65.9	76.7	85.6	79.0	51.8	73.7	56.9
CBS	82.1	70.9	69.1	55.6	90.1	62.6	74.2	71.0	62.4	52.2	87.6	56.9
RASS	78.3	87.9	81.2	62.7	82.3	**67.0**	75.6	86.5	78.4	54.6	77.3	**59.2**

(b) Asymmetric noise

Method	$\rho = 0.1$						$\rho = 0.05$					
	Asym-20%			Asym-40%			Asym-20%			Asym-40%		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Co-teaching	84.3	83.1	82.6	70.7	60.1	61.9	81.7	80.5	79.6	58.4	49.8	49.7
JoCoR	87.0	84.5	84.6	71.2	62.0	63.3	79.3	75.9	75.9	67.1	56.2	57.4
DivideMix	79.9	83.6	80.3	64.7	83.8	69.3	77.7	82.2	77.8	58.3	69.7	59.9
RoLT	86.1	81.3	82.0	73.8	76.6	72.2	82.3	79.5	79.1	72.1	74.5	70.1
TABASCO	89.3	83.3	84.2	76.5	77.0	73.6	86.4	80.3	81.1	74.2	72.0	69.7
CBS	93.1	34.9	46.8	77.7	58.4	62.7	88.4	38.4	47.4	76.0	59.2	61.5
RASS	93.5	80.2	**84.8**	83.6	75.4	**76.6**	91.6	78.1	**82.0**	80.6	70.3	**71.4**

L Computational Overhead

We analyzed the computational overhead of RASS on CIFAR-10 and CIFAR-100 under Sym-60% noise with an imbalance factor of $\rho = 0.05$. Table L1 compares the average time for training and sample selection per epoch with the baselines that address both class imbalance and noisy labels. Although RASS includes procedures for readiness evaluation and prototype-based sample selection, it showed competitive computational efficiency. For both CIFAR-10 and CIFAR-100, RASS required the lowest training time among the baselines. In terms of sample selection time, RASS required 0.670 and 0.880 seconds per epoch for CIFAR-10 and CIFAR-100, respectively. Although the number of classes increases from 10 to 100, the selection time increased

Table L1: Training and sample selection time per epoch on CIFAR-10 and CIFAR-100 under Sym-60% noise with $\rho = 0.05$.

Method	Training Time		Selection Time	
	CIFAR-10	CIFAR-100	CIFAR-10	CIFAR-100
TABASCO	9.039	7.810	0.702	1.200
RoLT	10.240	8.153	0.794	1.158
CBS	17.685	16.307	0.588	0.531
RASS	8.399	5.236	0.670	0.880

only by 0.210 seconds per epoch, indicating that the proposed selection procedure does not introduce substantial additional overhead as the number of classes increases.

Table L2 provides a detailed breakdown of the sample selection time of RASS. Most of the selection time was spent on the forward pass, which required 0.582 seconds on CIFAR-10 and 0.524 seconds on CIFAR-100. The time for GMM fitting increased from 0.078 seconds on CIFAR-10 to 0.307 seconds on CIFAR-100 as the number of classes increased. However, the forward pass still accounted for the largest portion of the selection time. In contrast, the overhead from readiness evaluation and prototype construction was negligible, requiring only 0.008 and 0.002 seconds on CIFAR-10, and 0.038 and 0.011 seconds on CIFAR-100, respectively. These results show that the additional cost introduced by the proposed readiness evaluation and prototype construction is small, while the main computational cost of sample selection comes from the forward pass and GMM fitting.

Table L2: Breakdown of the sample selection time of RASS per epoch on CIFAR-10 and CIFAR-100 under Sym-60% noise with $\rho = 0.05$.

Component	CIFAR-10	CIFAR-100
Forward pass	0.582	0.524
Readiness evaluation	0.008	0.038
Prototype construction	0.002	0.011
GMM fitting	0.078	0.307
Total	0.670	0.880

M Limitations

While RASS demonstrates strong performance across benchmark and real-world noisy datasets with class imbalance, its evaluation with pretrained and foundation models remains limited. Although we evaluate RASS using a CLIP-pretrained ViT, broader evaluations across different pretrained models and adaptation strategies are left for future work. Since such models often provide strong feature representations and may exhibit different memorization patterns from models trained from scratch, the readiness criterion may require further examination in this setting. Nevertheless, the core idea of delaying sample selection until class-specific features become sufficiently reliable can naturally extend to large-scale training scenarios where noisy and imbalanced annotations are common.

N Code Availability

The code for the proposed method is available at https://github.com/chihyeon-choi/RASS.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction clearly reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in Appendix M.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide the full assumptions and complete proof for all theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully discloses all the information and details to reproduce the main experimental results in Section 4.1 and Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the instructions, data, and code of the main experimental results in Section 4.1, Appendix E and N.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all the training and test details in Section 4.1 and Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report standard deviations in Tables 1, 2, 3, I1, I2, and H1. Statistical tests of the main results are given in Tables 1 and 2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the information on the computer resources in Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper conforms fully to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original papers for all datasets and comparison methods used in this study, and we follow their publicly available usage terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The paper used LLMs only for editing.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.