Supplementary Materials

S.1 Proof the Submodularity of Joint Entropy Function

Proof. To prove submodularity, we must show that for all $A \subseteq B \subseteq X$ and for any $s \in X \setminus B$, the following inequality holds:

$$H(A \cup \{s\}) - H(A) \geq H(B \cup \{s\}) - H(B).$$

This expression corresponds to the diminishing returns property. Recall that for entropy, the marginal gain of adding a random variable s to a set S can be expressed as the conditional entropy:

$$H(S \cup \{s\}) - H(S) = H(s \mid S),$$

where

$$H(s \mid S) = -\sum_{x_{S \cup \{s\}}} p(x_{S \cup \{s\}}) \log p(x_s \mid x_S).$$

Applying this identity, we have:

$$H(A \cup \{s\}) - H(A) = H(s \mid A), \quad H(B \cup \{s\}) - H(B) = H(s \mid B).$$

Since conditioning reduces entropy (or leaves it unchanged), and $A \subseteq B$, it follows that:

$$H(s \mid A) \geq H(s \mid B).$$

Therefore,

$$H(A \cup \{s\}) - H(A) \geq H(B \cup \{s\}) - H(B),$$

which establishes the submodularity of the entropy function. □

S.2 Proof of Proposition 1

We first state the following lemma:

Lemma S.1 (Proposition 4 in Gao and Pavel (2017))**.** *The softmax function is 1-Lipschitz with respect to the ℓ_2 norm; that is, for all $z, z' \in \mathbb{R}^p$,*

$$\|\sigma(z) - \sigma(z')\|_2 \leq \|z - z'\|_2,$$

where $\sigma(\cdot)$ denotes the softmax function.

Since $p(y = k \mid \boldsymbol{x}, \boldsymbol{\theta})$ is defined as the softmax function in our setting, and $z := \phi(\boldsymbol{x}^l_{k_1})$ and $z' := \phi(\boldsymbol{x}^u_{k_1})$ are assumed to be within a Euclidean distance of δ, we have

$$p(y = k_1 \mid \boldsymbol{x}^l_{k_1}, \boldsymbol{\theta}) - p(y = k_1 \mid \boldsymbol{x}^u_{k_1}, \boldsymbol{\theta}) \leq \|\phi(\boldsymbol{x}^l_{k_1}) - \phi(\boldsymbol{x}^u_{k_1})\|_2 \leq \delta,$$

which follows

$$p(y = k_1 \mid \boldsymbol{x}^l_{k_1}, \boldsymbol{\theta}) - \delta \leq p(y = k_1 \mid \boldsymbol{x}^u_{k_1}, \boldsymbol{\theta}).$$

Similarly, for class k_2,

$$-p(y = k_1 \mid \boldsymbol{x}^l_{k_2}, \boldsymbol{\theta}) - \delta \leq -p(y = k_1 \mid \boldsymbol{x}^u_{k_2}, \boldsymbol{\theta}).$$

Given the condition

$$p(y = k_1 \mid \boldsymbol{x}^l_{k_1}, \boldsymbol{\theta}) - p(y = k_1 \mid \boldsymbol{x}^l_{k_2}, \boldsymbol{\theta}) \geq 2\delta,$$

we obtain

$$p(y = k_1 \mid \boldsymbol{x}^u_{k_1}, \boldsymbol{\theta}) \geq p(y = k_1 \mid \boldsymbol{x}^u_{k_2}, \boldsymbol{\theta}).$$

Similarly, we also have

$$p(y = k_2 \mid \boldsymbol{x}^u_{k_2}, \boldsymbol{\theta}) \geq p(y = k_2 \mid \boldsymbol{x}^u_{k_1}, \boldsymbol{\theta}).$$

This concludes the proof.

S.3 Proof of Corollary 1

Fix an arbitrary class $k \in \{0, \ldots, K\}$, given an unlabeled sample $\boldsymbol{x}^u_k \in \boldsymbol{X}_k$ and its labeled counterpart $\boldsymbol{x}^l_k \in \boldsymbol{X}_k$, guaranteed by the assumption we have that

$$d\left(\phi(\boldsymbol{x}^u_k), \phi(\boldsymbol{x}^l_k)\right) \leq \delta.$$

From Lemma S.1 and the intermediate steps in the proof of Proposition 1, we have

$$p(y = k \mid \boldsymbol{x}^u_k, \boldsymbol{\theta}) \geq p(y = k \mid \boldsymbol{x}^l_k, \boldsymbol{\theta}) - \delta, \tag{S.1}$$

$$p(y = j \mid \boldsymbol{x}^u_k, \boldsymbol{\theta}) \leq p(y = j \mid \boldsymbol{x}^l_k, \boldsymbol{\theta}) + \delta, \tag{S.2}$$

for $j \neq k$. Therefore,

$$p(y = k \mid \boldsymbol{x}^u_k, \boldsymbol{\theta}) - p(y = j \mid \boldsymbol{x}^u_k, \boldsymbol{\theta}) \geq [p(y = k \mid \boldsymbol{x}^l_k, \boldsymbol{\theta}) - p(y = j \mid \boldsymbol{x}^l_k, \boldsymbol{\theta})] - 2\delta.$$

Since $p(y = j \mid \boldsymbol{x}^l_k, \boldsymbol{\theta}) \leq \max_{j' \neq k} p(y = j' \mid \boldsymbol{x}^l_k, \boldsymbol{\theta})$, the labeled-margin hypothesis yields, for every $j \neq k$,

$$p(y = k \mid \boldsymbol{x}^l_k, \boldsymbol{\theta}) - p(y = j \mid \boldsymbol{x}^l_k, \boldsymbol{\theta}) \;\geq\; p(y = k \mid \boldsymbol{x}^l_k, \boldsymbol{\theta}) - \max_{j' \neq k} p(y = j' \mid \boldsymbol{x}^l_k, \boldsymbol{\theta}) \;\geq\; 2\delta.$$

Combining these two equations, we have

$$p(y = k \mid \boldsymbol{x}^u_k, \boldsymbol{\theta}) - p(y = j \mid \boldsymbol{x}^u_k, \boldsymbol{\theta}) \;\geq\; 2\delta - 2\delta \;=\; 0 \qquad \text{for all } j \neq k,$$

i.e., $p(y = k \mid \boldsymbol{x}^u_k, \boldsymbol{\theta}) \geq p(y = j \mid \boldsymbol{x}^u_k, \boldsymbol{\theta})$ for all j. Therefore,

$$k \in \arg\max_j \; p(y = j \mid \boldsymbol{x}^u_k, \boldsymbol{\theta}).$$

Because k is arbitrarily chosen, this holds for every class $k \in \{0, \ldots, K\}$.

S.4 Proof of Proposition 2

We first prove the following lemma:

Lemma S.2. *Suppose* $p(y = k \mid \boldsymbol{x}, \boldsymbol{\theta}) = \frac{e^{h_k(\boldsymbol{x}|\boldsymbol{\theta}_k)}}{\sum_{k'} e^{h_{k'}(\boldsymbol{x}|\boldsymbol{\theta}_{k'})}}$ *is differentiable w.r.t.* $\boldsymbol{\theta}_k$ *and* $\boldsymbol{\theta}$ *lies in a bounded set* V*. Let* $\mathcal{X}$ *denote the set of samples observed in the online process, which is finite. Then*

$$\left\|\nabla_{\boldsymbol{\theta}} l^{(t+i)}(\boldsymbol{\theta})\right\|_2 \leq \sqrt{2}\, L,$$

where $L := \max_k \sup_{\boldsymbol{\theta} \in V,\, \boldsymbol{x} \in \mathcal{X}} \|\nabla_{\boldsymbol{\theta}} h_k(\boldsymbol{x}|\boldsymbol{\theta}_k)\|_2 < \infty$ *(by the finiteness of* $\mathcal{X}$*) is the upper bound on the logit gradients.*

Proof. It is well-known that

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{CE}}(\boldsymbol{x}, y \mid \boldsymbol{\theta}) = J_{\boldsymbol{\theta}}^{\top}(\boldsymbol{p} - \boldsymbol{e}_y),$$

where $\boldsymbol{p}$ is the predicted class probability vector, $\boldsymbol{e}_y$ is the one-hot encoding of the true label, and $J_{\boldsymbol{\theta}}$ is the Jacobian of the logits $\boldsymbol{z} = (h_k(\boldsymbol{x} \mid \boldsymbol{\theta}_k))_{k=1}^K$ with respect to $\boldsymbol{\theta}$.

Since $h_k(\boldsymbol{x} \mid \boldsymbol{\theta}_k)$ depends on $\boldsymbol{\theta}$ only through $\boldsymbol{\theta}_k$, the kth row of $J_{\boldsymbol{\theta}}$ equals $\boldsymbol{x}^{\top}$ in the block corresponding to $\boldsymbol{\theta}_k$ and $\mathbf{0}$ in the remaining blocks. The rows therefore occupy disjoint blocks of $\boldsymbol{\theta}$, which gives $J_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}^{\top} = \|\boldsymbol{x}\|_2^2 I$, and the Jacobian satisfies

$$\|J_{\boldsymbol{\theta}}\|_2 = \|\boldsymbol{x}\|_2 \leq L.$$

By the submultiplicative property of the operator norm,

$$\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{CE}}(\boldsymbol{x}, y \mid \boldsymbol{\theta})\|_2 \leq \|J_{\boldsymbol{\theta}}\|_2 \cdot \|\boldsymbol{p} - \boldsymbol{e}_y\|_2.$$

Since $\boldsymbol{p}$ lies in the probability simplex, we have $\|\boldsymbol{p}\|_2 \leq \|\boldsymbol{p}\|_1 = 1$ and $p_y \geq 0$. Therefore, $\|\boldsymbol{p} - \boldsymbol{e}_y\|_2^2 = \|\boldsymbol{p}\|_2^2 - 2p_y + 1 \leq 2$, that is, $\|\boldsymbol{p} - \boldsymbol{e}_y\|_2 \leq \sqrt{2}$. Therefore $\|\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{CE}}(\boldsymbol{x}, y \mid \boldsymbol{\theta})\|_2 \leq \sqrt{2}L$.

Since the ℓ_2-norm is convex, the norm of the average is upper-bounded by the average of the norms, hence

$$\left\|\nabla_{\boldsymbol{\theta}}\, l^{(t+i)}(\boldsymbol{\theta})\right\|_2 = \left\|\nabla_{\boldsymbol{\theta}} \left(\frac{1}{|D^{(t+i)}|} \sum_{(\boldsymbol{x},y) \in D^{(t+i)}} \mathcal{L}_{\mathrm{CE}}(\boldsymbol{x}, y \mid \boldsymbol{\theta}) \right) \right\|_2 \leq \sqrt{2}L.$$

□

The rest of the proof follows the (non-sublinear) regret analysis in Orabona (2019). For brevity, write $\boldsymbol{g}^{(t+i)} := \nabla_{\boldsymbol{\theta}}\, l^{(t+i)}(\boldsymbol{\theta}^{(t+i)})$, and let d denote the diameter of V. Since we consider softmax regression in this work, the logits are affine in $\boldsymbol{\theta}$, so each $\mathcal{L}_{\mathrm{CE}}(\boldsymbol{x}, y \mid \boldsymbol{\theta})$ is a log-sum-exp of an affine function minus a linear term, and is therefore convex in $\boldsymbol{\theta}$. Consequently $l^{(t+i)}$, an average of such losses, is convex, and we can use Lemma 2.12 in Orabona (2019)

to obtain

$$\begin{aligned}\sum_{i=1}^{T}\left(l^{(t+i)}(\boldsymbol{\theta}^{(t+i)})-l^{(t+i)}(\boldsymbol{\theta}_*^{(t)})\right) &\leq \sum_{i=1}^{T}\left(\frac{1}{2\eta^{(t+i)}}\left\|\boldsymbol{\theta}^{(t+i)}-\boldsymbol{\theta}_*^{(t)}\right\|_2^2-\frac{1}{2\eta^{(t+i)}}\left\|\boldsymbol{\theta}^{(t+i+1)}-\boldsymbol{\theta}_*^{(t)}\right\|_2^2\right)\\ &\quad+\frac{1}{2}\sum_{i=1}^{T}\eta^{(t+i)}\left\|\boldsymbol{g}^{(t+i)}\right\|_2^2,\\ &\leq \frac{1}{2\eta^{(t+1)}}d^2+d^2\sum_{i=1}^{T-1}\left(\frac{1}{2\eta^{(t+i+1)}}-\frac{1}{2\eta^{(t+i)}}\right)+\frac{1}{2}\sum_{i=1}^{T}\eta^{(t+i)}\left\|\boldsymbol{g}^{(t+i)}\right\|_2^2,\\ &= \frac{d^2}{2\eta^{(t+T)}}+\frac{1}{2}\sum_{i=1}^{T}\eta^{(t+i)}\left\|\boldsymbol{g}^{(t+i)}\right\|_2^2,\\ &\leq \frac{d^2}{2\eta^{(t+T)}}+\sum_{i=1}^{T}\eta^{(t+i)}L^2,\end{aligned}$$

where the last inequality uses Lemma S.2 ($\frac{1}{2}\|\boldsymbol{g}^{(t+i)}\|_2^2 \leq L^2$).

Since $\eta^{(t+i)}=\frac{1}{\sqrt{i}}$ (so $\eta^{(t+T)}=\frac{1}{\sqrt{T}}$), and

$$\sum_{i=1}^{T}\frac{1}{\sqrt{i}} \leq \int_1^T \frac{1}{\sqrt{i}}\,di+\frac{1}{\sqrt{1}}=\left[2\sqrt{i}\right]_1^T+1=2\sqrt{T}-2+1=2\sqrt{T}-1,$$

we obtain

$$\begin{aligned}\frac{d^2}{2\eta^{(t+T)}}+\sum_{i=1}^{T}\eta^{(t+i)}L^2 &\leq \frac{d^2}{2}\sqrt{T}+(2\sqrt{T}-1)L^2\\ &= O(\sqrt{T}).\end{aligned}$$

This completes the proof.

S.5 Comparison with Classical Strategies

In these experiments, we evaluate the performance of the proposed method by comparing it with the following widely used active learning strategies (Tharwat and Schenck 2023):

- **Random sampling (Random)**: This baseline strategy selects unlabeled samples uniformly at random.
- **Margin-based sampling (Margin)**: This method selects samples for which the difference between the top two predicted class probabilities is smallest.
- **Least confidence sampling (Confidence)**: This strategy queries the samples for which the model has the lowest maximum predicted class probability, reflecting the lowest confidence in classification.

- **Entropy-based sampling (Entropy)**: Selects samples with the highest predictive entropy.
- **Coreset sampling (Coreset)**: This approach aims to select a subset of samples that best represents the entire unlabeled dataset by maximizing coverage in the feature space measured by the Euclidean distance. To show the advantage of feature extractor selection, in these experiments, we use the raw features as the input of the coreset sampling strategy.

All the experiment settings are the same as Section 5.1. To ensure a fair comparison among the sampling strategies, the fault detector for each method is incrementally updated using the continual learning approach described in Section 4.3. Figure 1 and 2 show the results of the experiments when the fault occurrence probability $p_a = \{5\%, 10\%\}$, where the solid line indicates the mean metric across all replications, and the shaded region denotes the standard deviation.

From the results, we observe that (i) the proposed method consistently outperforms classical active learning strategies in terms of F1 score, precision, and recall across various settings. (ii) The performance gap between the proposed method and the classical strategies becomes more pronounced as the fault occurrence probability p_a decreases, indicating greater effectiveness of the proposed method under imbalanced conditions. (iii) Notably, the proposed method achieves a higher recall compared to classical strategies, with a larger margin than that observed in precision. This suggests that the proposed method is more effective at identifying minority fault types. (iv) Pure exploitation strategies (margin-based, least confidence-based, and entropy-based sampling) exhibit obvious performance degradation in the early stages. This is because they overlook sample correlations and the imbalance between majority and minority classes, thereby over-sampling from a specific region dominated by the majority class. In contrast, the proposed method effectively mitigates this exploitation bias.

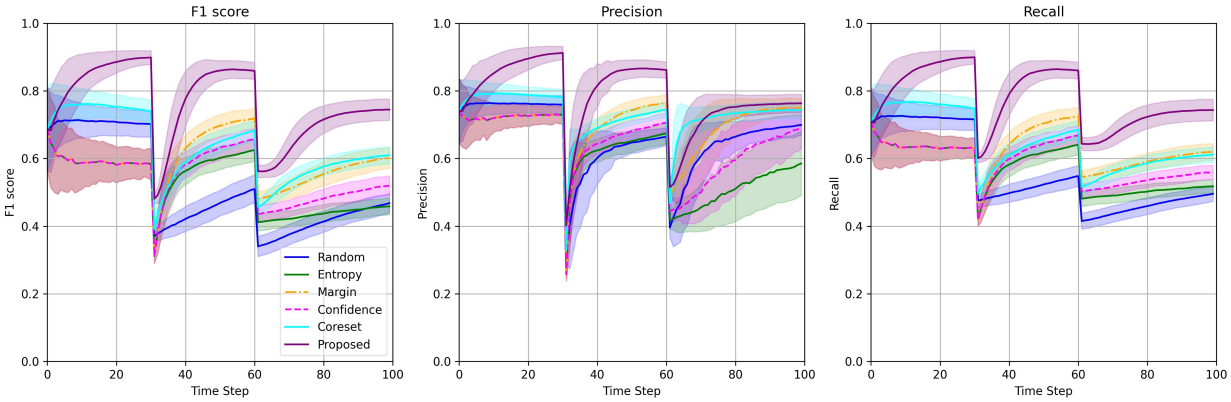
Therefore, the proposed method demonstrates superior performance in FDD, particularly in scenarios with limited samples of fault types.

S.6 Ablation Studies

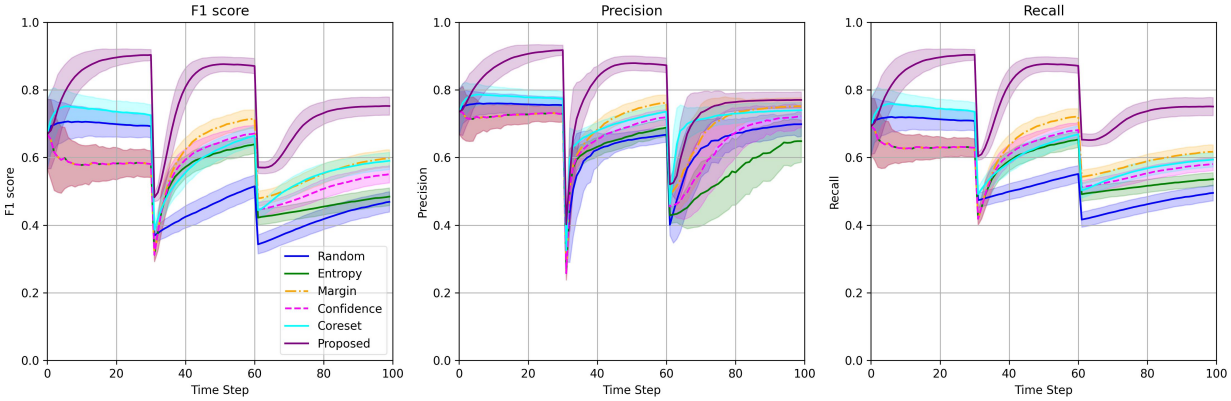
In this subsection, we conduct a series of ablation studies to evaluate the contribution of each part of the methodology.

We first analyze the detection performance of each component in the sampling strategy. In particular, we compare the precision, recall, and F1-score of the exploitation sampler (Exploit), the exploration sampler (Explore), the joint exploitation–exploration sampler (Joint E-E), which is the first-level sampler developed in Section 4.1, and the proposed method (Proposed), which is a bi-level sampler combining the joint exploitation–exploration sampler (Section 4.1) with the minority-aware replacement sampler (Section 4.2).

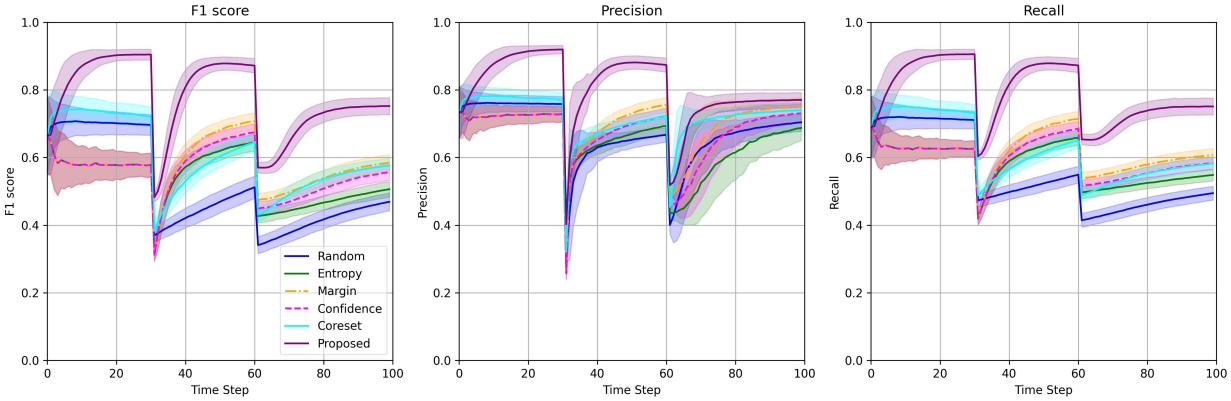
In the experiments, we consider the most difficult setting in Section 5.1 and Section 5.2, where the sampling rate is only 10% ($q = 10$) and the occurrence probability of each fault type is only 5% ($p_a = 5\%$). The replications are 500, and the results are shown in Figure 3,



(a) Results with the sampling rate of 10% ($q = 10$)

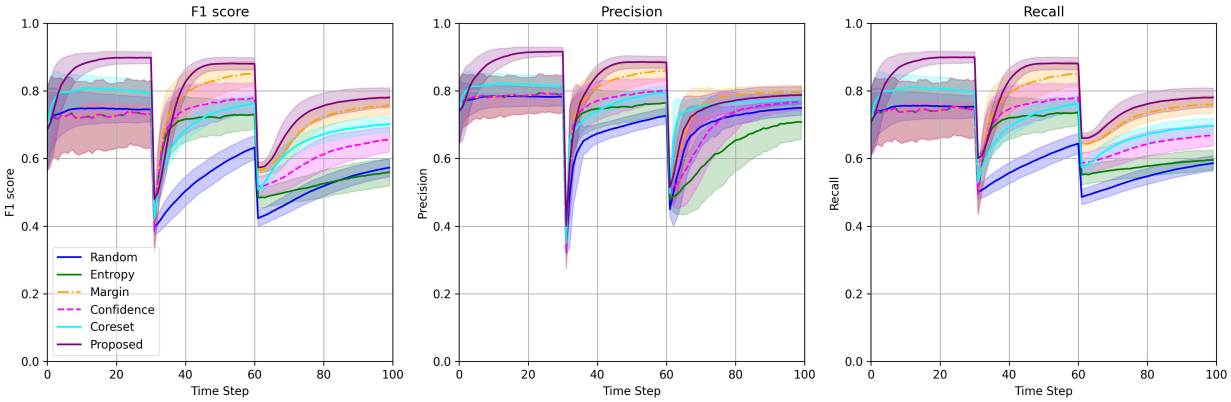


(b) Results with the sampling rate of 15% ($q = 15$)

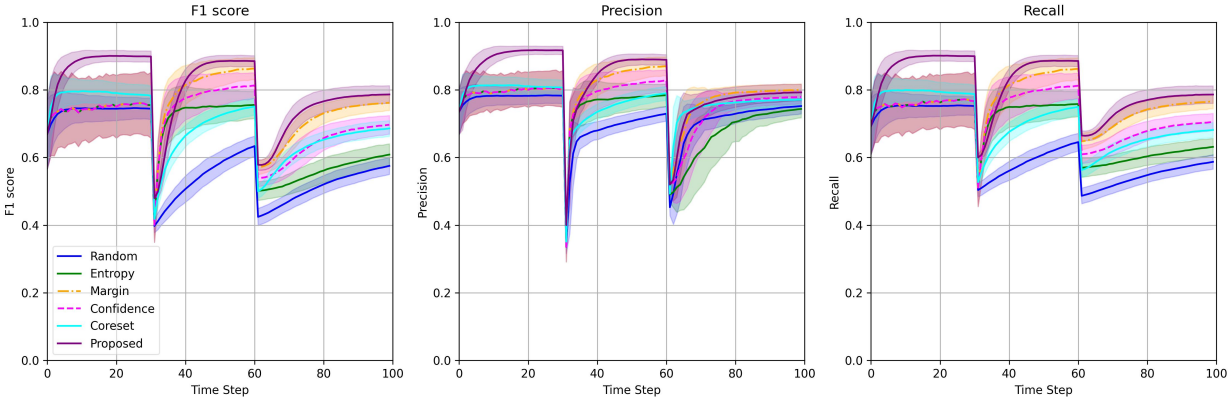


(c) Results with the sampling rate of 20% ($q = 20$)

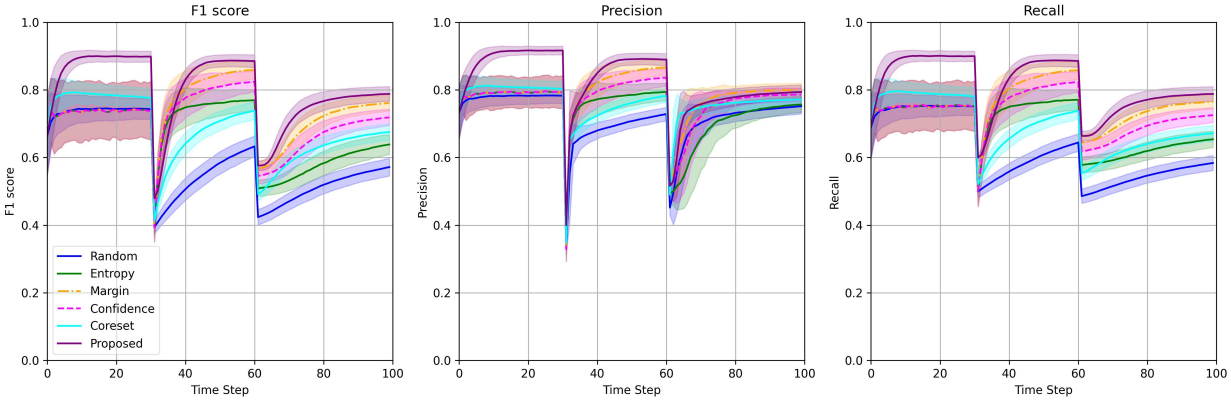
Figure 1: Comparison of performance metrics between the proposed method and classical active learning strategies under varying sampling rates, with $p_a = 5\%$, where the solid line indicates the mean metric and the shaded region denotes the standard deviation.



(a) Results with the sampling rate of 10% ($q = 10$)



(b) Results with the sampling rate of 15% ($q = 15$)



(c) Results with the sampling rate of 20% ($q = 20$)

Figure 2: Comparison of performance metrics between the proposed method and classical active learning strategies under varying sampling rates, with $p_a = 10\%$, where the solid line indicates the mean metric and the shaded region denotes the standard deviation.

where the solid line indicates the mean metric across all replications, and the shaded region denotes the standard deviation.

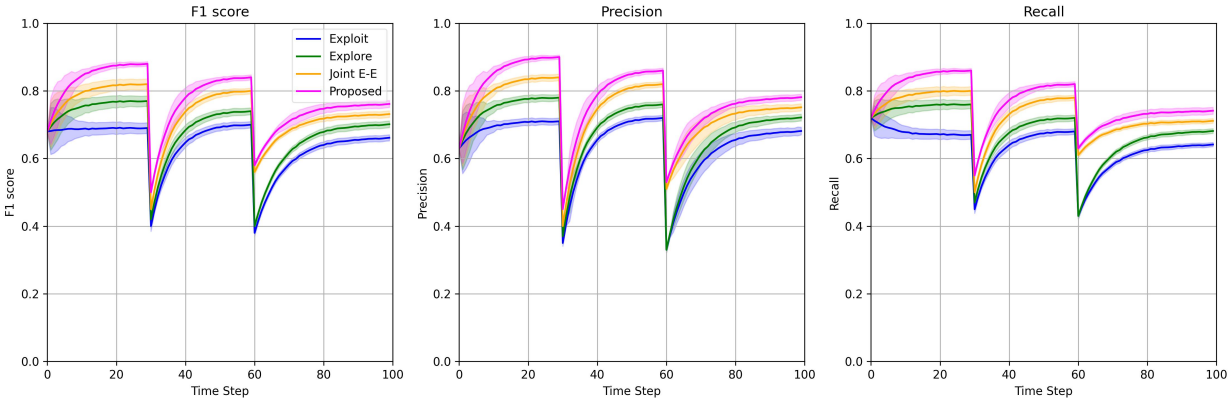


Figure 3: Comparison of performance metrics between different components of the sampling strategy with $p_a = 5\%$ and $q = 10$, where the solid line indicates the mean metric and the shaded region denotes the standard deviation.

As shown in Figure 3, the proposed bi-level sampler consistently outperforms the joint exploitation–exploration sampler, highlighting the critical role of the minority-aware replacement sampler in enhancing the informativeness of the training dataset. Furthermore, the joint exploitation–exploration sampler achieves better performance than the two pure strategies, that is, the exploitation and exploration samplers, demonstrating the effectiveness of balancing these two complementary approaches.

Notably, consistent with the results in Figure 2 in Section 5.2, the pure exploitation sampler exhibits the degradation in recall during the early phase, underscoring the necessity of exploration when data are limited and imbalanced, as exploration actively targets minority and previously unseen fault types. However, unlike classical exploitation strategies, the proposed exploitation strategy does not experience a similar drop in precision. This improvement stems from its ability to reduce redundancy in the training dataset, thereby mitigating the bias and improving the overall quality of the sampled data.

Next, we evaluate the performance of the continual learning method proposed in Section 4.3 by replacing it with retraining on the entire historical dataset (containing all labeled samples) within the same framework. The stopping criteria of the retraining method is either when the gradient falls below 10^{-3} or when the number of iterations reaches $10,000$.

In the experiments, we use the F1 score to evaluate the detection capability of the two methods and the running time per time step to assess their scalability and efficiency. All experiments are conducted on a MacBook equipped with an M1 Max chip and 32 GB of memory. Besides, we still set the sampling rate is only 10% ($q = 10$) and the occurrence probability of each fault type is only 5% ($p_a = 5\%$), which is a challenging case due to the low sampling rate and the sparse occurrence of faults. The replications are 200 and the results are shown in Figure 4.

As the figure shows, in each stage that starts by the arrival of a new fault type, the retraining method performs better than the continual learning method, which is demonstrated by a higher F1 score. However, this difference diminishes over time within each stage, as the

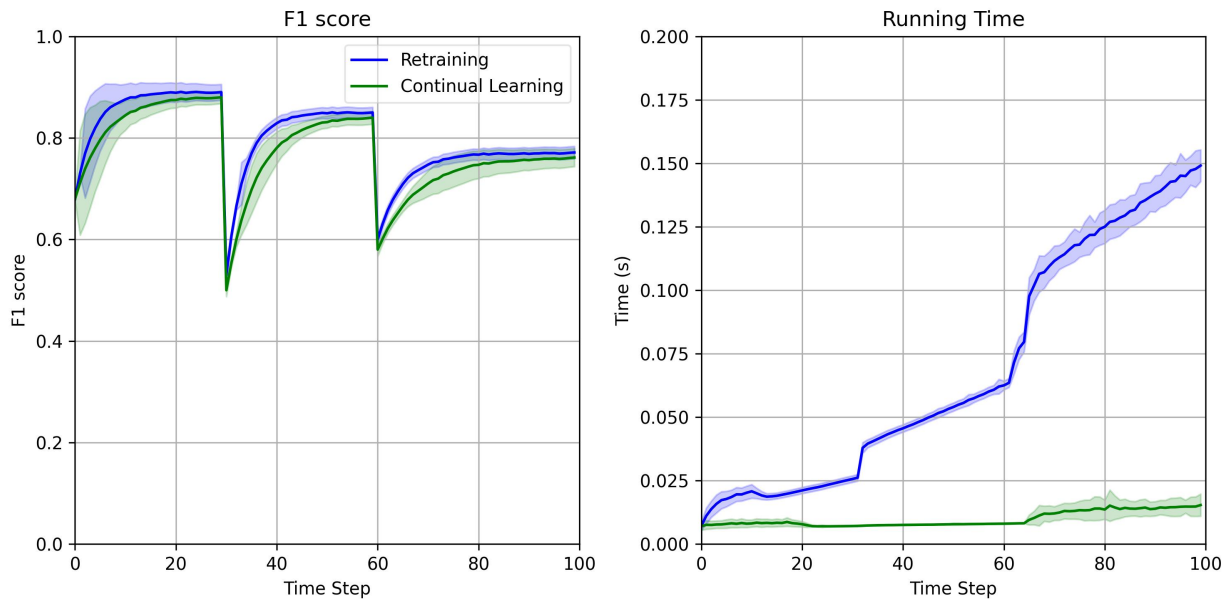


Figure 4: Comparison of performance metrics between the continual learning method and the retraining method with $p_a = 5\%$ and $q = 10$, where the solid line indicates the mean metric and the shaded region denotes the standard deviation.

parameters estimated by the continual learning method gradually converge to those output by the retraining method. In contrast, the computational cost (runtime) of retraining method increases rapidly over time, especially when a new fault type is introduced, whereas the continual learning approach maintains a nearly constant and much smaller runtime throughout the online process. These results demonstrate the scalability and efficiency of the continual learning method.

S.7 Robot Execution Faults

In this case study, a public dataset of robot execution failure is applied to demonstrate the effectiveness of the proposed method under a high-dimensional case (Lopes and Camarinha-Matos 1998) through comparisons with the benchmark methods introduced in Section 5.2. The dataset contains 90 features and records various fault types across five distinct stages: approach to grasp position, transfer of a part, position of part after a transfer failure, approach to ungrasp position, and motion with part. For this experiment, we focus on the approach to ungrasp position stage and consider collision and obstruction as the first and second fault types, respectively.

In these experiments, each method queries the labels of 10 samples from a batch of 100 unlabeled samples collected at each time step, i.e., the sampling rate is 10%. At the beginning, only 10 normal samples are available at $t = 0$ for pre-training the fault detector. The performance of each algorithm is evaluated using F1 score, precision, and recall, computed on a held-out balanced test set of size 200, which includes all fault types that may occur at the corresponding time step. Each experiment is replicated for 200 times.

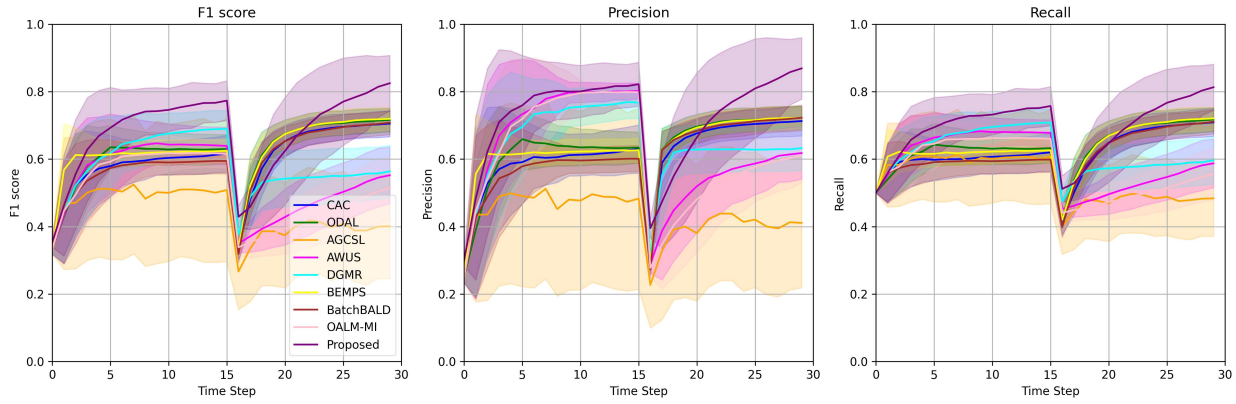


Figure 5: Comparison of performance metrics between the proposed method and benchmark algorithms in the Robot Execution Failure case study. The solid line represents the mean metric, and the shaded region indicates the standard deviation.

As described in Section 6.1, due to the limited number of original samples, we employ SMOTE to augment each fault type and enable the experiments for the online process. After data augmentation, collision and obstruction account for approximately 10% and 6% of the total samples, respectively. Each experiment is carried out over 30 time steps, with the collision fault type introduced at $t = 1$, and the obstruction fault type appearing from $t = 16$ onward.

The results are shown in Figure 5, where the proposed method outperforms all benchmarks across F1 score, precision, and recall throughout most of the time horizon. Interestingly, some benchmark methods exhibit slightly better performance at the very beginning of each stage when a new fault type is introduced. This is because their sampling strategies tend to focus on a specific region, which, in high-dimensional settings, can initially be beneficial—collected information tends to be informative, and redundancy has minimal impact early on. However, as more samples are acquired, the lack of diversity in the selected data begins to limit learning efficiency. In contrast, the proposed method demonstrates a faster and more sustained improvement, as it effectively accounts for redundancy and maintains a better balance between exploration and exploitation.

S.8 Quantitative Comparison with Benchmarks

In this section, we perform paired t-tests at representative time points where the performance of each method is stable ($t = 20$, $t = 50$, and $t = 90$) across 200 simulation replications to enable a clearer comparison between the proposed method and the benchmarks. The results are shown in the tables below.

The proposed method significantly outperforms all benchmark methods in terms of F1 score and recall under both $p_a = 5\%$ and $p_a = 10\%$ settings. For precision, the proposed method also achieves superior performance in almost all comparisons, with only two minor non-significant exceptions when $p_a = 10\%$.

Table 1: Paired t-test results for F1 score with $p_a = 5\%$ and the sampling rate of 10%.

t	Comparison	Proposed	Benchmark	Mean diff.	p-value
20	Proposed vs CAC	0.887 ± 0.026	0.739 ± 0.051	0.148	< 0.001
20	Proposed vs ODAL	0.887 ± 0.026	0.812 ± 0.036	0.075	< 0.001
20	Proposed vs AGCSL	0.887 ± 0.026	0.572 ± 0.073	0.315	< 0.001
20	Proposed vs AWUS	0.887 ± 0.026	0.705 ± 0.042	0.182	< 0.001
20	Proposed vs DGMR	0.887 ± 0.026	0.780 ± 0.042	0.107	< 0.001
20	Proposed vs BEMPS	0.887 ± 0.026	0.739 ± 0.034	0.148	< 0.001
20	Proposed vs BatchBALD	0.887 ± 0.026	0.736 ± 0.036	0.151	< 0.001
20	Proposed vs OALM-MI	0.887 ± 0.026	0.685 ± 0.047	0.202	< 0.001
50	Proposed vs CAC	0.866 ± 0.027	0.507 ± 0.052	0.359	< 0.001
50	Proposed vs ODAL	0.866 ± 0.027	0.654 ± 0.034	0.212	< 0.001
50	Proposed vs AGCSL	0.866 ± 0.027	0.442 ± 0.084	0.424	< 0.001
50	Proposed vs AWUS	0.866 ± 0.027	0.472 ± 0.043	0.394	< 0.001
50	Proposed vs DGMR	0.866 ± 0.027	0.612 ± 0.044	0.255	< 0.001
50	Proposed vs BEMPS	0.866 ± 0.027	0.507 ± 0.038	0.360	< 0.001
50	Proposed vs BatchBALD	0.866 ± 0.027	0.560 ± 0.043	0.306	< 0.001
50	Proposed vs OALM-MI	0.866 ± 0.027	0.546 ± 0.043	0.320	< 0.001
90	Proposed vs CAC	0.746 ± 0.033	0.501 ± 0.031	0.245	< 0.001
90	Proposed vs ODAL	0.746 ± 0.033	0.562 ± 0.030	0.184	< 0.001
90	Proposed vs AGCSL	0.746 ± 0.033	0.367 ± 0.064	0.379	< 0.001
90	Proposed vs AWUS	0.746 ± 0.033	0.450 ± 0.029	0.296	< 0.001
90	Proposed vs DGMR	0.746 ± 0.033	0.563 ± 0.030	0.183	< 0.001
90	Proposed vs BEMPS	0.746 ± 0.033	0.512 ± 0.028	0.234	< 0.001
90	Proposed vs BatchBALD	0.746 ± 0.033	0.574 ± 0.030	0.172	< 0.001
90	Proposed vs OALM-MI	0.746 ± 0.033	0.496 ± 0.031	0.250	< 0.001

Paired t-tests compare Proposed with each benchmark across matched random seeds; one-sided alternative: Proposed > Benchmark.

References

Gao B, Pavel L (2017) On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805* .

Lopes L, Camarinha-Matos L (1998) Robot Execution Failures. UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C5M89N.

Orabona F (2019) A modern introduction to online learning. *arXiv preprint arXiv:1912.13213* .

Tharwat A, Schenck W (2023) A survey on active learning: State-of-the-art, practical challenges and research directions. *Mathematics* 11(4):820.

Table 2: Paired t-test results for F1 score with $p_a = 5\%$ and the sampling rate of 15%.

t	Comparison	Proposed	Benchmark	Mean diff.	p-value
20	Proposed vs CAC	0.897 ± 0.020	0.724 ± 0.038	0.173	< 0.001
20	Proposed vs ODAL	0.897 ± 0.020	0.800 ± 0.030	0.097	< 0.001
20	Proposed vs AGCSL	0.897 ± 0.020	0.561 ± 0.072	0.335	< 0.001
20	Proposed vs AWUS	0.897 ± 0.020	0.707 ± 0.034	0.190	< 0.001
20	Proposed vs DGMR	0.897 ± 0.020	0.778 ± 0.038	0.119	< 0.001
20	Proposed vs BEMPS	0.897 ± 0.020	0.732 ± 0.029	0.165	< 0.001
20	Proposed vs BatchBALD	0.897 ± 0.020	0.738 ± 0.032	0.159	< 0.001
20	Proposed vs OALM-MI	0.897 ± 0.020	0.682 ± 0.046	0.215	< 0.001
50	Proposed vs CAC	0.878 ± 0.018	0.493 ± 0.042	0.385	< 0.001
50	Proposed vs ODAL	0.878 ± 0.018	0.629 ± 0.037	0.249	< 0.001
50	Proposed vs AGCSL	0.878 ± 0.018	0.431 ± 0.074	0.447	< 0.001
50	Proposed vs AWUS	0.878 ± 0.018	0.477 ± 0.040	0.401	< 0.001
50	Proposed vs DGMR	0.878 ± 0.018	0.544 ± 0.038	0.334	< 0.001
50	Proposed vs BEMPS	0.878 ± 0.018	0.518 ± 0.033	0.360	< 0.001
50	Proposed vs BatchBALD	0.878 ± 0.018	0.544 ± 0.041	0.334	< 0.001
50	Proposed vs OALM-MI	0.878 ± 0.018	0.597 ± 0.036	0.281	< 0.001
90	Proposed vs CAC	0.751 ± 0.029	0.500 ± 0.029	0.251	< 0.001
90	Proposed vs ODAL	0.751 ± 0.029	0.549 ± 0.025	0.202	< 0.001
90	Proposed vs AGCSL	0.751 ± 0.029	0.353 ± 0.055	0.398	< 0.001
90	Proposed vs AWUS	0.751 ± 0.029	0.452 ± 0.030	0.299	< 0.001
90	Proposed vs DGMR	0.751 ± 0.029	0.541 ± 0.028	0.210	< 0.001
90	Proposed vs BEMPS	0.751 ± 0.029	0.527 ± 0.028	0.224	< 0.001
90	Proposed vs BatchBALD	0.751 ± 0.029	0.553 ± 0.028	0.198	< 0.001
90	Proposed vs OALM-MI	0.751 ± 0.029	0.516 ± 0.027	0.235	< 0.001

Paired t-tests compare Proposed with each benchmark across matched random seeds; one-sided alternative: Proposed > Benchmark.

Table 3: Paired t-test results for F1 score with $p_a = 5\%$ and the sampling rate of 20%.

t	Comparison	Proposed	Benchmark	Mean diff.	p-value
20	Proposed vs CAC	0.898 ± 0.018	0.723 ± 0.030	0.175	< 0.001
20	Proposed vs ODAL	0.898 ± 0.018	0.792 ± 0.030	0.107	< 0.001
20	Proposed vs AGCSL	0.898 ± 0.018	0.545 ± 0.067	0.353	< 0.001
20	Proposed vs AWUS	0.898 ± 0.018	0.705 ± 0.032	0.193	< 0.001
20	Proposed vs DGMR	0.898 ± 0.018	0.780 ± 0.034	0.118	< 0.001
20	Proposed vs BEMPS	0.898 ± 0.018	0.733 ± 0.028	0.165	< 0.001
20	Proposed vs BatchBALD	0.898 ± 0.018	0.737 ± 0.032	0.161	< 0.001
20	Proposed vs OALM-MI	0.898 ± 0.018	0.663 ± 0.045	0.235	< 0.001
50	Proposed vs CAC	0.879 ± 0.019	0.484 ± 0.033	0.395	< 0.001
50	Proposed vs ODAL	0.879 ± 0.019	0.616 ± 0.033	0.263	< 0.001
50	Proposed vs AGCSL	0.879 ± 0.019	0.421 ± 0.065	0.458	< 0.001
50	Proposed vs AWUS	0.879 ± 0.019	0.473 ± 0.032	0.406	< 0.001
50	Proposed vs DGMR	0.879 ± 0.019	0.504 ± 0.035	0.375	< 0.001
50	Proposed vs BEMPS	0.879 ± 0.019	0.509 ± 0.034	0.370	< 0.001
50	Proposed vs BatchBALD	0.879 ± 0.019	0.539 ± 0.037	0.340	< 0.001
50	Proposed vs OALM-MI	0.879 ± 0.019	0.626 ± 0.035	0.253	< 0.001
90	Proposed vs CAC	0.751 ± 0.027	0.497 ± 0.024	0.255	< 0.001
90	Proposed vs ODAL	0.751 ± 0.027	0.540 ± 0.023	0.211	< 0.001
90	Proposed vs AGCSL	0.751 ± 0.027	0.343 ± 0.051	0.408	< 0.001
90	Proposed vs AWUS	0.751 ± 0.027	0.453 ± 0.027	0.298	< 0.001
90	Proposed vs DGMR	0.751 ± 0.027	0.514 ± 0.028	0.238	< 0.001
90	Proposed vs BEMPS	0.751 ± 0.027	0.529 ± 0.025	0.223	< 0.001
90	Proposed vs BatchBALD	0.751 ± 0.027	0.545 ± 0.029	0.206	< 0.001
90	Proposed vs OALM-MI	0.751 ± 0.027	0.529 ± 0.029	0.223	< 0.001

Paired t-tests compare Proposed with each benchmark across matched random seeds; one-sided alternative: Proposed $>$ Benchmark.

Table 4: Paired t-test results for Precision with $p_a = 5\%$ and the sampling rate of 10%.

t	Comparison	Proposed	Benchmark	Mean diff.	p-value
20	Proposed vs CAC	0.901 ± 0.026	0.778 ± 0.034	0.123	< 0.001
20	Proposed vs ODAL	0.901 ± 0.026	0.826 ± 0.027	0.075	< 0.001
20	Proposed vs AGCSL	0.901 ± 0.026	0.770 ± 0.023	0.131	< 0.001
20	Proposed vs AWUS	0.901 ± 0.026	0.760 ± 0.028	0.141	< 0.001
20	Proposed vs DGMR	0.901 ± 0.026	0.803 ± 0.030	0.098	< 0.001
20	Proposed vs BEMPS	0.901 ± 0.026	0.777 ± 0.025	0.124	< 0.001
20	Proposed vs BatchBALD	0.901 ± 0.026	0.774 ± 0.025	0.126	< 0.001
20	Proposed vs OALM-MI	0.901 ± 0.026	0.754 ± 0.028	0.147	< 0.001
50	Proposed vs CAC	0.870 ± 0.025	0.659 ± 0.029	0.211	< 0.001
50	Proposed vs ODAL	0.870 ± 0.025	0.730 ± 0.027	0.141	< 0.001
50	Proposed vs AGCSL	0.870 ± 0.025	0.741 ± 0.103	0.129	< 0.001
50	Proposed vs AWUS	0.870 ± 0.025	0.644 ± 0.034	0.226	< 0.001
50	Proposed vs DGMR	0.870 ± 0.025	0.724 ± 0.024	0.147	< 0.001
50	Proposed vs BEMPS	0.870 ± 0.025	0.668 ± 0.024	0.202	< 0.001
50	Proposed vs BatchBALD	0.870 ± 0.025	0.687 ± 0.026	0.184	< 0.001
50	Proposed vs OALM-MI	0.870 ± 0.025	0.673 ± 0.028	0.198	< 0.001
90	Proposed vs CAC	0.765 ± 0.028	0.701 ± 0.071	0.064	< 0.001
90	Proposed vs ODAL	0.765 ± 0.028	0.737 ± 0.036	0.028	< 0.001
90	Proposed vs AGCSL	0.765 ± 0.028	0.706 ± 0.096	0.059	< 0.001
90	Proposed vs AWUS	0.765 ± 0.028	0.689 ± 0.031	0.076	< 0.001
90	Proposed vs DGMR	0.765 ± 0.028	0.754 ± 0.024	0.012	< 0.001
90	Proposed vs BEMPS	0.765 ± 0.028	0.716 ± 0.025	0.050	< 0.001
90	Proposed vs BatchBALD	0.765 ± 0.028	0.727 ± 0.024	0.038	< 0.001
90	Proposed vs OALM-MI	0.765 ± 0.028	0.718 ± 0.030	0.048	< 0.001

Paired t-tests compare Proposed with each benchmark across matched random seeds; one-sided alternative: Proposed > Benchmark.

Table 5: Paired t-test results for Precision with $p_a = 5\%$ and the sampling rate of 15%.

t	Comparison	Proposed	Benchmark	Mean diff.	p-value
20	Proposed vs CAC	0.910 ± 0.020	0.769 ± 0.026	0.141	< 0.001
20	Proposed vs ODAL	0.910 ± 0.020	0.818 ± 0.022	0.093	< 0.001
20	Proposed vs AGCSL	0.910 ± 0.020	0.771 ± 0.022	0.139	< 0.001
20	Proposed vs AWUS	0.910 ± 0.020	0.762 ± 0.024	0.148	< 0.001
20	Proposed vs DGMR	0.910 ± 0.020	0.800 ± 0.028	0.110	< 0.001
20	Proposed vs BEMPS	0.910 ± 0.020	0.773 ± 0.021	0.137	< 0.001
20	Proposed vs BatchBALD	0.910 ± 0.020	0.775 ± 0.025	0.135	< 0.001
20	Proposed vs OALM-MI	0.910 ± 0.020	0.754 ± 0.028	0.157	< 0.001
50	Proposed vs CAC	0.882 ± 0.018	0.657 ± 0.030	0.225	< 0.001
50	Proposed vs ODAL	0.882 ± 0.018	0.721 ± 0.025	0.160	< 0.001
50	Proposed vs AGCSL	0.882 ± 0.018	0.754 ± 0.085	0.128	< 0.001
50	Proposed vs AWUS	0.882 ± 0.018	0.653 ± 0.029	0.228	< 0.001
50	Proposed vs DGMR	0.882 ± 0.018	0.702 ± 0.025	0.179	< 0.001
50	Proposed vs BEMPS	0.882 ± 0.018	0.676 ± 0.022	0.206	< 0.001
50	Proposed vs BatchBALD	0.882 ± 0.018	0.682 ± 0.024	0.199	< 0.001
50	Proposed vs OALM-MI	0.882 ± 0.018	0.694 ± 0.024	0.188	< 0.001
90	Proposed vs CAC	0.769 ± 0.025	0.710 ± 0.056	0.060	< 0.001
90	Proposed vs ODAL	0.769 ± 0.025	0.739 ± 0.033	0.031	< 0.001
90	Proposed vs AGCSL	0.769 ± 0.025	0.725 ± 0.089	0.045	< 0.001
90	Proposed vs AWUS	0.769 ± 0.025	0.694 ± 0.031	0.075	< 0.001
90	Proposed vs DGMR	0.769 ± 0.025	0.748 ± 0.023	0.021	< 0.001
90	Proposed vs BEMPS	0.769 ± 0.025	0.722 ± 0.024	0.047	< 0.001
90	Proposed vs BatchBALD	0.769 ± 0.025	0.723 ± 0.024	0.046	< 0.001
90	Proposed vs OALM-MI	0.769 ± 0.025	0.731 ± 0.035	0.039	< 0.001

Paired t-tests compare Proposed with each benchmark across matched random seeds; one-sided alternative: Proposed $>$ Benchmark.

Table 6: Paired t-test results for Precision with $p_a = 5\%$ and the sampling rate of 20%.

t	Comparison	Proposed	Benchmark	Mean diff.	p-value
20	Proposed vs CAC	0.912 ± 0.017	0.770 ± 0.022	0.142	< 0.001
20	Proposed vs ODAL	0.912 ± 0.017	0.813 ± 0.023	0.099	< 0.001
20	Proposed vs AGCSL	0.912 ± 0.017	0.767 ± 0.024	0.145	< 0.001
20	Proposed vs AWUS	0.912 ± 0.017	0.761 ± 0.022	0.151	< 0.001
20	Proposed vs DGMR	0.912 ± 0.017	0.801 ± 0.026	0.111	< 0.001
20	Proposed vs BEMPS	0.912 ± 0.017	0.774 ± 0.021	0.138	< 0.001
20	Proposed vs BatchBALD	0.912 ± 0.017	0.775 ± 0.024	0.137	< 0.001
20	Proposed vs OALM-MI	0.912 ± 0.017	0.747 ± 0.028	0.165	< 0.001
50	Proposed vs CAC	0.883 ± 0.019	0.658 ± 0.024	0.225	< 0.001
50	Proposed vs ODAL	0.883 ± 0.019	0.714 ± 0.024	0.169	< 0.001
50	Proposed vs AGCSL	0.883 ± 0.019	0.771 ± 0.046	0.111	< 0.001
50	Proposed vs AWUS	0.883 ± 0.019	0.656 ± 0.025	0.227	< 0.001
50	Proposed vs DGMR	0.883 ± 0.019	0.690 ± 0.027	0.192	< 0.001
50	Proposed vs BEMPS	0.883 ± 0.019	0.672 ± 0.021	0.211	< 0.001
50	Proposed vs BatchBALD	0.883 ± 0.019	0.683 ± 0.022	0.200	< 0.001
50	Proposed vs OALM-MI	0.883 ± 0.019	0.703 ± 0.027	0.180	< 0.001
90	Proposed vs CAC	0.769 ± 0.023	0.720 ± 0.036	0.049	< 0.001
90	Proposed vs ODAL	0.769 ± 0.023	0.735 ± 0.029	0.034	< 0.001
90	Proposed vs AGCSL	0.769 ± 0.023	0.746 ± 0.066	0.023	< 0.001
90	Proposed vs AWUS	0.769 ± 0.023	0.696 ± 0.030	0.073	< 0.001
90	Proposed vs DGMR	0.769 ± 0.023	0.738 ± 0.023	0.031	< 0.001
90	Proposed vs BEMPS	0.769 ± 0.023	0.724 ± 0.023	0.045	< 0.001
90	Proposed vs BatchBALD	0.769 ± 0.023	0.724 ± 0.023	0.045	< 0.001
90	Proposed vs OALM-MI	0.769 ± 0.023	0.735 ± 0.032	0.034	< 0.001

Paired t-tests compare Proposed with each benchmark across matched random seeds; one-sided alternative: Proposed $>$ Benchmark.

Table 7: Paired t-test results for Recall with $p_a = 5\%$ and the sampling rate of 10%.

t	Comparison	Proposed	Benchmark	Mean diff.	p-value
20	Proposed vs CAC	0.888 ± 0.025	0.747 ± 0.045	0.140	< 0.001
20	Proposed vs ODAL	0.888 ± 0.025	0.814 ± 0.034	0.074	< 0.001
20	Proposed vs AGCSL	0.888 ± 0.025	0.631 ± 0.048	0.257	< 0.001
20	Proposed vs AWUS	0.888 ± 0.025	0.718 ± 0.035	0.170	< 0.001
20	Proposed vs DGMR	0.888 ± 0.025	0.785 ± 0.038	0.103	< 0.001
20	Proposed vs BEMPS	0.888 ± 0.025	0.747 ± 0.030	0.141	< 0.001
20	Proposed vs BatchBALD	0.888 ± 0.025	0.744 ± 0.032	0.144	< 0.001
20	Proposed vs OALM-MI	0.888 ± 0.025	0.703 ± 0.038	0.185	< 0.001
50	Proposed vs CAC	0.867 ± 0.026	0.548 ± 0.037	0.319	< 0.001
50	Proposed vs ODAL	0.867 ± 0.026	0.668 ± 0.029	0.199	< 0.001
50	Proposed vs AGCSL	0.867 ± 0.026	0.495 ± 0.057	0.372	< 0.001
50	Proposed vs AWUS	0.867 ± 0.026	0.525 ± 0.030	0.342	< 0.001
50	Proposed vs DGMR	0.867 ± 0.026	0.633 ± 0.035	0.234	< 0.001
50	Proposed vs BEMPS	0.867 ± 0.026	0.549 ± 0.027	0.318	< 0.001
50	Proposed vs BatchBALD	0.867 ± 0.026	0.585 ± 0.034	0.282	< 0.001
50	Proposed vs OALM-MI	0.867 ± 0.026	0.572 ± 0.034	0.295	< 0.001
90	Proposed vs CAC	0.747 ± 0.032	0.547 ± 0.022	0.200	< 0.001
90	Proposed vs ODAL	0.747 ± 0.032	0.586 ± 0.024	0.160	< 0.001
90	Proposed vs AGCSL	0.747 ± 0.032	0.417 ± 0.048	0.330	< 0.001
90	Proposed vs AWUS	0.747 ± 0.032	0.483 ± 0.022	0.264	< 0.001
90	Proposed vs DGMR	0.747 ± 0.032	0.582 ± 0.024	0.165	< 0.001
90	Proposed vs BEMPS	0.747 ± 0.032	0.537 ± 0.022	0.210	< 0.001
90	Proposed vs BatchBALD	0.747 ± 0.032	0.581 ± 0.026	0.166	< 0.001
90	Proposed vs OALM-MI	0.747 ± 0.032	0.522 ± 0.026	0.225	< 0.001

Paired t-tests compare Proposed with each benchmark across matched random seeds; one-sided alternative: Proposed > Benchmark.

Table 8: Paired t-test results for Recall with $p_a = 5\%$ and the sampling rate of 15%.

t	Comparison	Proposed	Benchmark	Mean diff.	p-value
20	Proposed vs CAC	0.898 ± 0.020	0.735 ± 0.032	0.163	< 0.001
20	Proposed vs ODAL	0.898 ± 0.020	0.803 ± 0.028	0.094	< 0.001
20	Proposed vs AGCSL	0.898 ± 0.020	0.625 ± 0.047	0.273	< 0.001
20	Proposed vs AWUS	0.898 ± 0.020	0.720 ± 0.028	0.177	< 0.001
20	Proposed vs DGMR	0.898 ± 0.020	0.782 ± 0.034	0.115	< 0.001
20	Proposed vs BEMPS	0.898 ± 0.020	0.741 ± 0.025	0.156	< 0.001
20	Proposed vs BatchBALD	0.898 ± 0.020	0.745 ± 0.029	0.152	< 0.001
20	Proposed vs OALM-MI	0.898 ± 0.020	0.700 ± 0.038	0.198	< 0.001
50	Proposed vs CAC	0.878 ± 0.018	0.540 ± 0.030	0.339	< 0.001
50	Proposed vs ODAL	0.878 ± 0.018	0.648 ± 0.030	0.230	< 0.001
50	Proposed vs AGCSL	0.878 ± 0.018	0.487 ± 0.049	0.392	< 0.001
50	Proposed vs AWUS	0.878 ± 0.018	0.529 ± 0.030	0.349	< 0.001
50	Proposed vs DGMR	0.878 ± 0.018	0.589 ± 0.029	0.289	< 0.001
50	Proposed vs BEMPS	0.878 ± 0.018	0.558 ± 0.025	0.321	< 0.001
50	Proposed vs BatchBALD	0.878 ± 0.018	0.574 ± 0.031	0.305	< 0.001
50	Proposed vs OALM-MI	0.878 ± 0.018	0.611 ± 0.030	0.267	< 0.001
90	Proposed vs CAC	0.751 ± 0.028	0.540 ± 0.021	0.211	< 0.001
90	Proposed vs ODAL	0.751 ± 0.028	0.574 ± 0.021	0.177	< 0.001
90	Proposed vs AGCSL	0.751 ± 0.028	0.404 ± 0.040	0.348	< 0.001
90	Proposed vs AWUS	0.751 ± 0.028	0.485 ± 0.024	0.267	< 0.001
90	Proposed vs DGMR	0.751 ± 0.028	0.557 ± 0.023	0.194	< 0.001
90	Proposed vs BEMPS	0.751 ± 0.028	0.547 ± 0.022	0.204	< 0.001
90	Proposed vs BatchBALD	0.751 ± 0.028	0.563 ± 0.023	0.188	< 0.001
90	Proposed vs OALM-MI	0.751 ± 0.028	0.543 ± 0.022	0.208	< 0.001

Paired t-tests compare Proposed with each benchmark across matched random seeds; one-sided alternative: Proposed > Benchmark.

Table 9: Paired t-test results for Recall with $p_a = 5\%$ and the sampling rate of 20%.

t	Comparison	Proposed	Benchmark	Mean diff.	p-value
20	Proposed vs CAC	0.899 ± 0.018	0.733 ± 0.026	0.166	< 0.001
20	Proposed vs ODAL	0.899 ± 0.018	0.795 ± 0.028	0.104	< 0.001
20	Proposed vs AGCSL	0.899 ± 0.018	0.614 ± 0.043	0.285	< 0.001
20	Proposed vs AWUS	0.899 ± 0.018	0.718 ± 0.028	0.181	< 0.001
20	Proposed vs DGMR	0.899 ± 0.018	0.784 ± 0.032	0.115	< 0.001
20	Proposed vs BEMPS	0.899 ± 0.018	0.742 ± 0.025	0.158	< 0.001
20	Proposed vs BatchBALD	0.899 ± 0.018	0.745 ± 0.028	0.154	< 0.001
20	Proposed vs OALM-MI	0.899 ± 0.018	0.686 ± 0.035	0.213	< 0.001
50	Proposed vs CAC	0.880 ± 0.019	0.534 ± 0.023	0.345	< 0.001
50	Proposed vs ODAL	0.880 ± 0.019	0.637 ± 0.026	0.243	< 0.001
50	Proposed vs AGCSL	0.880 ± 0.019	0.479 ± 0.043	0.401	< 0.001
50	Proposed vs AWUS	0.880 ± 0.019	0.526 ± 0.023	0.354	< 0.001
50	Proposed vs DGMR	0.880 ± 0.019	0.565 ± 0.026	0.315	< 0.001
50	Proposed vs BEMPS	0.880 ± 0.019	0.551 ± 0.025	0.329	< 0.001
50	Proposed vs BatchBALD	0.880 ± 0.019	0.571 ± 0.029	0.308	< 0.001
50	Proposed vs OALM-MI	0.880 ± 0.019	0.635 ± 0.031	0.245	< 0.001
90	Proposed vs CAC	0.752 ± 0.027	0.534 ± 0.019	0.218	< 0.001
90	Proposed vs ODAL	0.752 ± 0.027	0.566 ± 0.020	0.186	< 0.001
90	Proposed vs AGCSL	0.752 ± 0.027	0.395 ± 0.035	0.357	< 0.001
90	Proposed vs AWUS	0.752 ± 0.027	0.484 ± 0.021	0.268	< 0.001
90	Proposed vs DGMR	0.752 ± 0.027	0.533 ± 0.023	0.218	< 0.001
90	Proposed vs BEMPS	0.752 ± 0.027	0.547 ± 0.020	0.205	< 0.001
90	Proposed vs BatchBALD	0.752 ± 0.027	0.556 ± 0.024	0.196	< 0.001
90	Proposed vs OALM-MI	0.752 ± 0.027	0.558 ± 0.021	0.193	< 0.001

Paired t-tests compare Proposed with each benchmark across matched random seeds; one-sided alternative: Proposed > Benchmark.

Table 10: Paired t-test results for F1 score with $p_a = 10\%$ and the sampling rate of 10%.

t	Comparison	Proposed	Benchmark	Mean diff.	p-value
20	Proposed vs CAC	0.889 ± 0.019	0.751 ± 0.043	0.138	< 0.001
20	Proposed vs ODAL	0.889 ± 0.019	0.848 ± 0.028	0.041	< 0.001
20	Proposed vs AGCSL	0.889 ± 0.019	0.674 ± 0.063	0.215	< 0.001
20	Proposed vs AWUS	0.889 ± 0.019	0.751 ± 0.040	0.138	< 0.001
20	Proposed vs DGMR	0.889 ± 0.019	0.835 ± 0.038	0.053	< 0.001
20	Proposed vs BEMPS	0.889 ± 0.019	0.772 ± 0.033	0.116	< 0.001
20	Proposed vs BatchBALD	0.889 ± 0.019	0.775 ± 0.035	0.114	< 0.001
20	Proposed vs OALM-MI	0.889 ± 0.019	0.750 ± 0.046	0.139	< 0.001
50	Proposed vs CAC	0.880 ± 0.019	0.639 ± 0.042	0.241	< 0.001
50	Proposed vs ODAL	0.880 ± 0.019	0.751 ± 0.030	0.129	< 0.001
50	Proposed vs AGCSL	0.880 ± 0.019	0.611 ± 0.070	0.270	< 0.001
50	Proposed vs AWUS	0.880 ± 0.019	0.584 ± 0.043	0.296	< 0.001
50	Proposed vs DGMR	0.880 ± 0.019	0.744 ± 0.034	0.137	< 0.001
50	Proposed vs BEMPS	0.880 ± 0.019	0.617 ± 0.031	0.263	< 0.001
50	Proposed vs BatchBALD	0.880 ± 0.019	0.673 ± 0.034	0.207	< 0.001
50	Proposed vs OALM-MI	0.880 ± 0.019	0.665 ± 0.039	0.215	< 0.001
90	Proposed vs CAC	0.775 ± 0.032	0.605 ± 0.037	0.170	< 0.001
90	Proposed vs ODAL	0.775 ± 0.032	0.659 ± 0.030	0.116	< 0.001
90	Proposed vs AGCSL	0.775 ± 0.032	0.516 ± 0.059	0.259	< 0.001
90	Proposed vs AWUS	0.775 ± 0.032	0.556 ± 0.027	0.219	< 0.001
90	Proposed vs DGMR	0.775 ± 0.032	0.680 ± 0.029	0.095	< 0.001
90	Proposed vs BEMPS	0.775 ± 0.032	0.621 ± 0.027	0.154	< 0.001
90	Proposed vs BatchBALD	0.775 ± 0.032	0.687 ± 0.026	0.088	< 0.001
90	Proposed vs OALM-MI	0.775 ± 0.032	0.606 ± 0.027	0.169	< 0.001

Paired t-tests compare Proposed with each benchmark across matched random seeds; one-sided alternative: Proposed > Benchmark.

Table 11: Paired t-test results for F1 score with $p_a = 10\%$ and the sampling rate of 15%.

t	Comparison	Proposed	Benchmark	Mean diff.	p-value
20	Proposed vs CAC	0.893 ± 0.017	0.745 ± 0.034	0.149	< 0.001
20	Proposed vs ODAL	0.893 ± 0.017	0.834 ± 0.027	0.060	< 0.001
20	Proposed vs AGCSL	0.893 ± 0.017	0.658 ± 0.070	0.235	< 0.001
20	Proposed vs AWUS	0.893 ± 0.017	0.748 ± 0.032	0.145	< 0.001
20	Proposed vs DGMR	0.893 ± 0.017	0.819 ± 0.030	0.074	< 0.001
20	Proposed vs BEMPS	0.893 ± 0.017	0.772 ± 0.030	0.122	< 0.001
20	Proposed vs BatchBALD	0.893 ± 0.017	0.773 ± 0.030	0.121	< 0.001
20	Proposed vs OALM-MI	0.893 ± 0.017	0.774 ± 0.038	0.120	< 0.001
50	Proposed vs CAC	0.884 ± 0.018	0.620 ± 0.034	0.264	< 0.001
50	Proposed vs ODAL	0.884 ± 0.018	0.734 ± 0.029	0.150	< 0.001
50	Proposed vs AGCSL	0.884 ± 0.018	0.583 ± 0.064	0.301	< 0.001
50	Proposed vs AWUS	0.884 ± 0.018	0.587 ± 0.036	0.297	< 0.001
50	Proposed vs DGMR	0.884 ± 0.018	0.674 ± 0.034	0.210	< 0.001
50	Proposed vs BEMPS	0.884 ± 0.018	0.630 ± 0.034	0.255	< 0.001
50	Proposed vs BatchBALD	0.884 ± 0.018	0.655 ± 0.036	0.229	< 0.001
50	Proposed vs OALM-MI	0.884 ± 0.018	0.719 ± 0.031	0.166	< 0.001
90	Proposed vs CAC	0.780 ± 0.028	0.599 ± 0.029	0.181	< 0.001
90	Proposed vs ODAL	0.780 ± 0.028	0.646 ± 0.024	0.134	< 0.001
90	Proposed vs AGCSL	0.780 ± 0.028	0.485 ± 0.056	0.296	< 0.001
90	Proposed vs AWUS	0.780 ± 0.028	0.557 ± 0.028	0.223	< 0.001
90	Proposed vs DGMR	0.780 ± 0.028	0.646 ± 0.026	0.134	< 0.001
90	Proposed vs BEMPS	0.780 ± 0.028	0.634 ± 0.025	0.146	< 0.001
90	Proposed vs BatchBALD	0.780 ± 0.028	0.671 ± 0.025	0.109	< 0.001
90	Proposed vs OALM-MI	0.780 ± 0.028	0.632 ± 0.024	0.149	< 0.001

Paired t-tests compare Proposed with each benchmark across matched random seeds; one-sided alternative: Proposed > Benchmark.

Table 12: Paired t-test results for F1 score with $p_a = 10\%$ and the sampling rate of 20%.

t	Comparison	Proposed	Benchmark	Mean diff.	p-value
20	Proposed vs CAC	0.892 ± 0.018	0.743 ± 0.034	0.148	< 0.001
20	Proposed vs ODAL	0.892 ± 0.018	0.832 ± 0.024	0.060	< 0.001
20	Proposed vs AGCSL	0.892 ± 0.018	0.653 ± 0.067	0.239	< 0.001
20	Proposed vs AWUS	0.892 ± 0.018	0.745 ± 0.029	0.147	< 0.001
20	Proposed vs DGMR	0.892 ± 0.018	0.809 ± 0.032	0.083	< 0.001
20	Proposed vs BEMPS	0.892 ± 0.018	0.773 ± 0.026	0.119	< 0.001
20	Proposed vs BatchBALD	0.892 ± 0.018	0.774 ± 0.029	0.118	< 0.001
20	Proposed vs OALM-MI	0.892 ± 0.018	0.781 ± 0.042	0.111	< 0.001
50	Proposed vs CAC	0.888 ± 0.017	0.608 ± 0.035	0.280	< 0.001
50	Proposed vs ODAL	0.888 ± 0.017	0.722 ± 0.029	0.165	< 0.001
50	Proposed vs AGCSL	0.888 ± 0.017	0.552 ± 0.069	0.335	< 0.001
50	Proposed vs AWUS	0.888 ± 0.017	0.585 ± 0.037	0.303	< 0.001
50	Proposed vs DGMR	0.888 ± 0.017	0.621 ± 0.034	0.267	< 0.001
50	Proposed vs BEMPS	0.888 ± 0.017	0.620 ± 0.031	0.268	< 0.001
50	Proposed vs BatchBALD	0.888 ± 0.017	0.648 ± 0.030	0.239	< 0.001
50	Proposed vs OALM-MI	0.888 ± 0.017	0.766 ± 0.027	0.122	< 0.001
90	Proposed vs CAC	0.784 ± 0.025	0.593 ± 0.029	0.191	< 0.001
90	Proposed vs ODAL	0.784 ± 0.025	0.644 ± 0.027	0.140	< 0.001
90	Proposed vs AGCSL	0.784 ± 0.025	0.458 ± 0.053	0.326	< 0.001
90	Proposed vs AWUS	0.784 ± 0.025	0.556 ± 0.025	0.228	< 0.001
90	Proposed vs DGMR	0.784 ± 0.025	0.611 ± 0.027	0.173	< 0.001
90	Proposed vs BEMPS	0.784 ± 0.025	0.634 ± 0.026	0.151	< 0.001
90	Proposed vs BatchBALD	0.784 ± 0.025	0.658 ± 0.026	0.127	< 0.001
90	Proposed vs OALM-MI	0.784 ± 0.025	0.663 ± 0.027	0.121	< 0.001

Paired t-tests compare Proposed with each benchmark across matched random seeds; one-sided alternative: Proposed > Benchmark.

Table 13: Paired t-test results for Precision with $p_a = 10\%$ and the sampling rate of 10%.

t	Comparison	Proposed	Benchmark	Mean diff.	p-value
20	Proposed vs CAC	0.909 ± 0.015	0.785 ± 0.029	0.124	< 0.001
20	Proposed vs ODAL	0.909 ± 0.015	0.852 ± 0.024	0.056	< 0.001
20	Proposed vs AGCSL	0.909 ± 0.015	0.796 ± 0.022	0.113	< 0.001
20	Proposed vs AWUS	0.909 ± 0.015	0.786 ± 0.028	0.123	< 0.001
20	Proposed vs DGMR	0.909 ± 0.015	0.842 ± 0.031	0.066	< 0.001
20	Proposed vs BEMPS	0.909 ± 0.015	0.797 ± 0.025	0.112	< 0.001
20	Proposed vs BatchBALD	0.909 ± 0.015	0.798 ± 0.026	0.111	< 0.001
20	Proposed vs OALM-MI	0.909 ± 0.015	0.788 ± 0.030	0.121	< 0.001
50	Proposed vs CAC	0.886 ± 0.019	0.712 ± 0.028	0.174	< 0.001
50	Proposed vs ODAL	0.886 ± 0.019	0.791 ± 0.022	0.095	< 0.001
50	Proposed vs AGCSL	0.886 ± 0.019	0.795 ± 0.025	0.091	< 0.001
50	Proposed vs AWUS	0.886 ± 0.019	0.705 ± 0.025	0.180	< 0.001
50	Proposed vs DGMR	0.886 ± 0.019	0.796 ± 0.021	0.090	< 0.001
50	Proposed vs BEMPS	0.886 ± 0.019	0.725 ± 0.022	0.160	< 0.001
50	Proposed vs BatchBALD	0.886 ± 0.019	0.749 ± 0.023	0.137	< 0.001
50	Proposed vs OALM-MI	0.886 ± 0.019	0.740 ± 0.026	0.146	< 0.001
90	Proposed vs CAC	0.783 ± 0.030	0.748 ± 0.032	0.034	< 0.001
90	Proposed vs ODAL	0.783 ± 0.030	0.780 ± 0.019	0.003	0.091
90	Proposed vs AGCSL	0.783 ± 0.030	0.760 ± 0.057	0.023	< 0.001
90	Proposed vs AWUS	0.783 ± 0.030	0.745 ± 0.023	0.038	< 0.001
90	Proposed vs DGMR	0.783 ± 0.030	0.785 ± 0.019	-0.002	0.880
90	Proposed vs BEMPS	0.783 ± 0.030	0.757 ± 0.020	0.026	< 0.001
90	Proposed vs BatchBALD	0.783 ± 0.030	0.767 ± 0.018	0.016	< 0.001
90	Proposed vs OALM-MI	0.783 ± 0.030	0.761 ± 0.021	0.022	< 0.001

Paired t-tests compare Proposed with each benchmark across matched random seeds; one-sided alternative: Proposed > Benchmark.

Table 14: Paired t-test results for Precision with $p_a = 10\%$ and the sampling rate of 15%.

t	Comparison	Proposed	Benchmark	Mean diff.	p-value
20	Proposed vs CAC	0.913 ± 0.013	0.781 ± 0.025	0.131	< 0.001
20	Proposed vs ODAL	0.913 ± 0.013	0.841 ± 0.023	0.072	< 0.001
20	Proposed vs AGCSL	0.913 ± 0.013	0.792 ± 0.023	0.120	< 0.001
20	Proposed vs AWUS	0.913 ± 0.013	0.785 ± 0.022	0.128	< 0.001
20	Proposed vs DGMR	0.913 ± 0.013	0.829 ± 0.024	0.084	< 0.001
20	Proposed vs BEMPS	0.913 ± 0.013	0.797 ± 0.023	0.116	< 0.001
20	Proposed vs BatchBALD	0.913 ± 0.013	0.796 ± 0.025	0.117	< 0.001
20	Proposed vs OALM-MI	0.913 ± 0.013	0.802 ± 0.027	0.111	< 0.001
50	Proposed vs CAC	0.890 ± 0.018	0.706 ± 0.026	0.184	< 0.001
50	Proposed vs ODAL	0.890 ± 0.018	0.783 ± 0.021	0.107	< 0.001
50	Proposed vs AGCSL	0.890 ± 0.018	0.794 ± 0.019	0.096	< 0.001
50	Proposed vs AWUS	0.890 ± 0.018	0.709 ± 0.022	0.180	< 0.001
50	Proposed vs DGMR	0.890 ± 0.018	0.765 ± 0.020	0.124	< 0.001
50	Proposed vs BEMPS	0.890 ± 0.018	0.733 ± 0.022	0.157	< 0.001
50	Proposed vs BatchBALD	0.890 ± 0.018	0.742 ± 0.022	0.148	< 0.001
50	Proposed vs OALM-MI	0.890 ± 0.018	0.773 ± 0.022	0.117	< 0.001
90	Proposed vs CAC	0.787 ± 0.026	0.752 ± 0.024	0.035	< 0.001
90	Proposed vs ODAL	0.787 ± 0.026	0.778 ± 0.019	0.009	< 0.001
90	Proposed vs AGCSL	0.787 ± 0.026	0.767 ± 0.047	0.020	< 0.001
90	Proposed vs AWUS	0.787 ± 0.026	0.745 ± 0.025	0.042	< 0.001
90	Proposed vs DGMR	0.787 ± 0.026	0.778 ± 0.018	0.009	< 0.001
90	Proposed vs BEMPS	0.787 ± 0.026	0.762 ± 0.019	0.025	< 0.001
90	Proposed vs BatchBALD	0.787 ± 0.026	0.764 ± 0.019	0.023	< 0.001
90	Proposed vs OALM-MI	0.787 ± 0.026	0.773 ± 0.021	0.014	< 0.001

Paired t-tests compare Proposed with each benchmark across matched random seeds; one-sided alternative: Proposed > Benchmark.

Table 15: Paired t-test results for Precision with $p_a = 10\%$ and the sampling rate of 20%.

t	Comparison	Proposed	Benchmark	Mean diff.	p-value
20	Proposed vs CAC	0.912 ± 0.013	0.781 ± 0.025	0.131	< 0.001
20	Proposed vs ODAL	0.912 ± 0.013	0.840 ± 0.021	0.072	< 0.001
20	Proposed vs AGCSL	0.912 ± 0.013	0.792 ± 0.022	0.120	< 0.001
20	Proposed vs AWUS	0.912 ± 0.013	0.783 ± 0.021	0.128	< 0.001
20	Proposed vs DGMR	0.912 ± 0.013	0.821 ± 0.026	0.091	< 0.001
20	Proposed vs BEMPS	0.912 ± 0.013	0.798 ± 0.021	0.114	< 0.001
20	Proposed vs BatchBALD	0.912 ± 0.013	0.797 ± 0.023	0.115	< 0.001
20	Proposed vs OALM-MI	0.912 ± 0.013	0.807 ± 0.029	0.105	< 0.001
50	Proposed vs CAC	0.893 ± 0.017	0.702 ± 0.024	0.191	< 0.001
50	Proposed vs ODAL	0.893 ± 0.017	0.775 ± 0.020	0.119	< 0.001
50	Proposed vs AGCSL	0.893 ± 0.017	0.791 ± 0.019	0.102	< 0.001
50	Proposed vs AWUS	0.893 ± 0.017	0.708 ± 0.022	0.185	< 0.001
50	Proposed vs DGMR	0.893 ± 0.017	0.743 ± 0.020	0.150	< 0.001
50	Proposed vs BEMPS	0.893 ± 0.017	0.729 ± 0.020	0.164	< 0.001
50	Proposed vs BatchBALD	0.893 ± 0.017	0.740 ± 0.019	0.153	< 0.001
50	Proposed vs OALM-MI	0.893 ± 0.017	0.801 ± 0.021	0.092	< 0.001
90	Proposed vs CAC	0.790 ± 0.023	0.751 ± 0.024	0.039	< 0.001
90	Proposed vs ODAL	0.790 ± 0.023	0.776 ± 0.021	0.014	< 0.001
90	Proposed vs AGCSL	0.790 ± 0.023	0.770 ± 0.037	0.020	< 0.001
90	Proposed vs AWUS	0.790 ± 0.023	0.746 ± 0.021	0.044	< 0.001
90	Proposed vs DGMR	0.790 ± 0.023	0.771 ± 0.019	0.019	< 0.001
90	Proposed vs BEMPS	0.790 ± 0.023	0.762 ± 0.020	0.028	< 0.001
90	Proposed vs BatchBALD	0.790 ± 0.023	0.763 ± 0.020	0.027	< 0.001
90	Proposed vs OALM-MI	0.790 ± 0.023	0.782 ± 0.021	0.008	< 0.001

Paired t-tests compare Proposed with each benchmark across matched random seeds; one-sided alternative: Proposed > Benchmark.

Table 16: Paired t-test results for Recall with $p_a = 10\%$ and the sampling rate of 10%.

t	Comparison	Proposed	Benchmark	Mean diff.	p-value
20	Proposed vs CAC	0.890 ± 0.019	0.758 ± 0.038	0.132	< 0.001
20	Proposed vs ODAL	0.890 ± 0.019	0.848 ± 0.027	0.041	< 0.001
20	Proposed vs AGCSL	0.890 ± 0.019	0.702 ± 0.047	0.187	< 0.001
20	Proposed vs AWUS	0.890 ± 0.019	0.758 ± 0.036	0.132	< 0.001
20	Proposed vs DGMR	0.890 ± 0.019	0.837 ± 0.036	0.053	< 0.001
20	Proposed vs BEMPS	0.890 ± 0.019	0.777 ± 0.031	0.113	< 0.001
20	Proposed vs BatchBALD	0.890 ± 0.019	0.779 ± 0.032	0.111	< 0.001
20	Proposed vs OALM-MI	0.890 ± 0.019	0.757 ± 0.041	0.132	< 0.001
50	Proposed vs CAC	0.881 ± 0.019	0.648 ± 0.037	0.233	< 0.001
50	Proposed vs ODAL	0.881 ± 0.019	0.754 ± 0.027	0.127	< 0.001
50	Proposed vs AGCSL	0.881 ± 0.019	0.622 ± 0.059	0.259	< 0.001
50	Proposed vs AWUS	0.881 ± 0.019	0.608 ± 0.033	0.273	< 0.001
50	Proposed vs DGMR	0.881 ± 0.019	0.749 ± 0.030	0.132	< 0.001
50	Proposed vs BEMPS	0.881 ± 0.019	0.637 ± 0.027	0.244	< 0.001
50	Proposed vs BatchBALD	0.881 ± 0.019	0.681 ± 0.030	0.201	< 0.001
50	Proposed vs OALM-MI	0.881 ± 0.019	0.672 ± 0.034	0.209	< 0.001
90	Proposed vs CAC	0.777 ± 0.031	0.633 ± 0.026	0.144	< 0.001
90	Proposed vs ODAL	0.777 ± 0.031	0.672 ± 0.025	0.105	< 0.001
90	Proposed vs AGCSL	0.777 ± 0.031	0.536 ± 0.048	0.241	< 0.001
90	Proposed vs AWUS	0.777 ± 0.031	0.574 ± 0.022	0.203	< 0.001
90	Proposed vs DGMR	0.777 ± 0.031	0.692 ± 0.024	0.085	< 0.001
90	Proposed vs BEMPS	0.777 ± 0.031	0.635 ± 0.022	0.142	< 0.001
90	Proposed vs BatchBALD	0.777 ± 0.031	0.687 ± 0.024	0.090	< 0.001
90	Proposed vs OALM-MI	0.777 ± 0.031	0.619 ± 0.023	0.158	< 0.001

Paired t-tests compare Proposed with each benchmark across matched random seeds; one-sided alternative: Proposed > Benchmark.

Table 17: Paired t-test results for Recall with $p_a = 10\%$ and the sampling rate of 15%.

t	Comparison	Proposed	Benchmark	Mean diff.	p-value
20	Proposed vs CAC	0.894 ± 0.016	0.752 ± 0.030	0.142	< 0.001
20	Proposed vs ODAL	0.894 ± 0.016	0.835 ± 0.027	0.059	< 0.001
20	Proposed vs AGCSL	0.894 ± 0.016	0.692 ± 0.052	0.203	< 0.001
20	Proposed vs AWUS	0.894 ± 0.016	0.756 ± 0.028	0.139	< 0.001
20	Proposed vs DGMR	0.894 ± 0.016	0.821 ± 0.029	0.074	< 0.001
20	Proposed vs BEMPS	0.894 ± 0.016	0.777 ± 0.027	0.118	< 0.001
20	Proposed vs BatchBALD	0.894 ± 0.016	0.776 ± 0.028	0.118	< 0.001
20	Proposed vs OALM-MI	0.894 ± 0.016	0.779 ± 0.034	0.116	< 0.001
50	Proposed vs CAC	0.885 ± 0.018	0.633 ± 0.031	0.252	< 0.001
50	Proposed vs ODAL	0.885 ± 0.018	0.739 ± 0.027	0.146	< 0.001
50	Proposed vs AGCSL	0.885 ± 0.018	0.599 ± 0.052	0.286	< 0.001
50	Proposed vs AWUS	0.885 ± 0.018	0.611 ± 0.029	0.274	< 0.001
50	Proposed vs DGMR	0.885 ± 0.018	0.692 ± 0.028	0.193	< 0.001
50	Proposed vs BEMPS	0.885 ± 0.018	0.648 ± 0.027	0.237	< 0.001
50	Proposed vs BatchBALD	0.885 ± 0.018	0.666 ± 0.031	0.219	< 0.001
50	Proposed vs OALM-MI	0.885 ± 0.018	0.721 ± 0.029	0.164	< 0.001
90	Proposed vs CAC	0.782 ± 0.028	0.625 ± 0.022	0.157	< 0.001
90	Proposed vs ODAL	0.782 ± 0.028	0.661 ± 0.021	0.121	< 0.001
90	Proposed vs AGCSL	0.782 ± 0.028	0.506 ± 0.044	0.276	< 0.001
90	Proposed vs AWUS	0.782 ± 0.028	0.575 ± 0.024	0.207	< 0.001
90	Proposed vs DGMR	0.782 ± 0.028	0.654 ± 0.023	0.128	< 0.001
90	Proposed vs BEMPS	0.782 ± 0.028	0.644 ± 0.021	0.138	< 0.001
90	Proposed vs BatchBALD	0.782 ± 0.028	0.671 ± 0.024	0.111	< 0.001
90	Proposed vs OALM-MI	0.782 ± 0.028	0.645 ± 0.021	0.137	< 0.001

Paired t-tests compare Proposed with each benchmark across matched random seeds; one-sided alternative: Proposed > Benchmark.

Table 18: Paired t-test results for Recall with $p_a = 10\%$ and the sampling rate of 20%.

t	Comparison	Proposed	Benchmark	Mean diff.	p-value
20	Proposed vs CAC	0.893 ± 0.017	0.751 ± 0.030	0.142	< 0.001
20	Proposed vs ODAL	0.893 ± 0.017	0.833 ± 0.024	0.060	< 0.001
20	Proposed vs AGCSL	0.893 ± 0.017	0.687 ± 0.050	0.206	< 0.001
20	Proposed vs AWUS	0.893 ± 0.017	0.753 ± 0.026	0.141	< 0.001
20	Proposed vs DGMR	0.893 ± 0.017	0.811 ± 0.031	0.082	< 0.001
20	Proposed vs BEMPS	0.893 ± 0.017	0.777 ± 0.025	0.116	< 0.001
20	Proposed vs BatchBALD	0.893 ± 0.017	0.778 ± 0.027	0.115	< 0.001
20	Proposed vs OALM-MI	0.893 ± 0.017	0.786 ± 0.038	0.107	< 0.001
50	Proposed vs CAC	0.889 ± 0.017	0.623 ± 0.029	0.265	< 0.001
50	Proposed vs ODAL	0.889 ± 0.017	0.728 ± 0.026	0.161	< 0.001
50	Proposed vs AGCSL	0.889 ± 0.017	0.575 ± 0.055	0.313	< 0.001
50	Proposed vs AWUS	0.889 ± 0.017	0.609 ± 0.029	0.280	< 0.001
50	Proposed vs DGMR	0.889 ± 0.017	0.650 ± 0.027	0.239	< 0.001
50	Proposed vs BEMPS	0.889 ± 0.017	0.640 ± 0.026	0.248	< 0.001
50	Proposed vs BatchBALD	0.889 ± 0.017	0.660 ± 0.026	0.228	< 0.001
50	Proposed vs OALM-MI	0.889 ± 0.017	0.765 ± 0.026	0.123	< 0.001
90	Proposed vs CAC	0.785 ± 0.024	0.619 ± 0.022	0.167	< 0.001
90	Proposed vs ODAL	0.785 ± 0.024	0.657 ± 0.023	0.128	< 0.001
90	Proposed vs AGCSL	0.785 ± 0.024	0.484 ± 0.042	0.301	< 0.001
90	Proposed vs AWUS	0.785 ± 0.024	0.573 ± 0.021	0.213	< 0.001
90	Proposed vs DGMR	0.785 ± 0.024	0.619 ± 0.023	0.167	< 0.001
90	Proposed vs BEMPS	0.785 ± 0.024	0.643 ± 0.023	0.142	< 0.001
90	Proposed vs BatchBALD	0.785 ± 0.024	0.659 ± 0.023	0.126	< 0.001
90	Proposed vs OALM-MI	0.785 ± 0.024	0.675 ± 0.023	0.110	< 0.001

Paired t-tests compare Proposed with each benchmark across matched random seeds; one-sided alternative: Proposed > Benchmark.