TOWARDS RELIABLE UNCERTAINTY QUANTIFICATION: WHEN CONFORMAL PREDICTION MEETS CLASS INCREMENTAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable uncertainty quantification is critical for Class-Incremental Learning (CIL), making Conformal Prediction (CP) a promising approach due to its finite-sample coverage guarantees. However, under a fixed calibration budget, CP must allocate limited calibration samples across classes, while discarded samples from old classes cannot be recovered. As class-wise calibration demands evolve over time, allocations tailored to current CP reliability may become inadequate for future tasks. To address these challenges, we propose **BACA**, a **B**udget-**A**ware **C**alibration **A**llocation framework that improves CP reliability in CIL scenarios. Specifically, BACA contains three complementary components: (1) ***Current Demand Estimation*** estimates class-wise calibration demand by characterizing the sensitivity of prediction-set instability to calibration size; (2) ***Irreversibility-Aware Future Protection*** identifies classes vulnerable to irreversible future risk and protects them from excessive calibration-sample removal; and (3) ***Marginal Budget Allocation***, which reallocates the fixed calibration budget by balancing the marginal stability benefit of adding samples and the marginal instability cost of removing samples. Extensive experiments on three benchmarks with four CIL learners show that BACA consistently improves prediction-set stability under fixed calibration budgets. It achieves higher Jaccard similarity and lower disagreement than uniform allocation while maintaining comparable coverage and average prediction-set size.

1 INTRODUCTION

Class-Incremental Learning (CIL) enables a model to continuously acquire new classes while preserving previously learned knowledge (Rebuffi et al., 2017; Masana et al., 2023). For incrementally updated models deployed in real-world applications, predictive accuracy alone is insufficient; predictive uncertainty should also be quantified. Conformal Prediction (CP) provides a principled approach to uncertainty quantification with finite-sample coverage guarantees under exchangeability (Angelopoulos & Bates, 2023). However, reliable uncertainty quantification requires more than valid coverage and compact prediction sets. It is also desirable that prediction sets remain stable across different finite calibration subsets.

This stability is important because, in practical deployment, different finite calibration samples may produce different prediction sets for the same model and input. These differences can change downstream decisions and make uncertainty quantification less trustworthy Bian & Barber (2023); Papamichalis et al. (2026); Kladny et al. (2026). This issue becomes particularly important in CIL under a fixed calibration budget. As new classes arrive, the same calibration memory must be shared among an increasing number of classes, while discarded old-class calibration samples cannot be recovered when future calibration demand increases.

As shown in Fig. 1, prediction-set disagreement varies substantially across classes on both CIFAR-100 and Tiny-ImageNet, even with the same class-wise calibration size. Meanwhile, the disagreement of some classes can increase as incremental learning proceeds, such that a class with low current disagreement may become more calibration-demanding in later tasks. These observations reveal two key challenges for calibration allocation: *(i)* how to characterize the current calibration

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

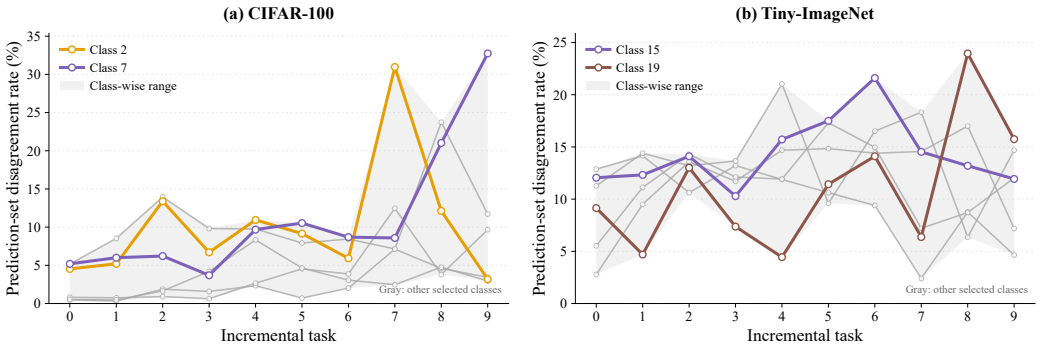


Figure 1: Class-wise prediction-set disagreement on CIFAR-100 and Tiny-ImageNet with $\alpha = 0.1$ and $n = 20$ calibration samples per class. Six first-task classes are evenly selected; shaded regions show the class-wise range, and colored trajectories highlight representative classes with increasing later-task disagreement.

demand of each class and allocate the limited samples accordingly, and *(ii)* how to account for future demand changes when discarded old-class calibration samples cannot be recovered.

To address these challenges, we propose **BACA** (**B**udget-**A**ware **C**alibration **A**llocation), a framework that dynamically allocates limited calibration samples across classes to support reliable uncertainty quantification in CIL. The key idea of BACA is to allocate calibration memory by jointly considering the current reliability requirement of each class and the risk caused by irreversible calibration-sample deletion. BACA first introduces ***Current Demand Estimation***, which characterizes how prediction-set instability changes with calibration size and derives class-wise calibration demand. By capturing the sensitivity of prediction-set stability to calibration resources, this stage enables the calibration budget to be allocated according to the current uncertainty quantification requirement of different classes.

However, directly allocating calibration memory according to current demand may discard samples from classes that become more important in later incremental stages. To address this issue, BACA introduces ***Irreversibility-Aware Future Protection***, which identifies classes vulnerable to future calibration insufficiency and adjusts their allocation priorities before irreversible deletion occurs. Based on the protected calibration demand, BACA further performs ***Marginal Budget Allocation***, which optimizes the fixed calibration memory by transferring samples among classes according to the marginal stability benefit of adding samples and the marginal instability cost of removing samples. This design enables BACA to maintain reliable uncertainty quantification throughout the incremental learning process under limited calibration memory.

We evaluate BACA on three benchmarks with four CIL learners under various calibration budgets and miscoverage levels. Extensive experiments demonstrate that BACA consistently improves prediction-set stability compared with uniform allocation, achieving higher Jaccard similarity and lower disagreement while maintaining comparable empirical coverage and average prediction-set size. Moreover, robustness experiments under different calibration constraints show the effectiveness of BACA, and further analyses provide insights into the effectiveness of current demand estimation and irreversibility-aware future protection. Our main contributions are summarized as follows:

- **A New Problem**. We reveal the calibration allocation problem in CIL under fixed and irreversible calibration memory, where class-wise demands are heterogeneous and evolve.
- **A Novel Framework**. We propose **BACA**, a Budget-Aware Calibration Allocation framework that dynamically allocates calibration samples by considering current reliability demand and irreversible future risk for adaptive calibration management.
- **Superior Performance**. Extensive experiments on three benchmarks with four CIL learners demonstrate that BACA improves prediction-set stability while preserving comparable coverage and set size.

2 Related Work

Class-Incremental Learning (CIL) aims to enable models to learn new classes sequentially while retaining previously acquired knowledge without task identities during inference Rebuffi et al. (2017); Masana et al. (2023); Zhou et al. (2024). Existing approaches mainly mitigate catastrophic forgetting through replay, regularization, knowledge distillation, and parameter isolation Kirkpatrick et al. (2017); Lopez-Paz & Ranzato (2017); Buzzega et al. (2020). Among them, replay-based methods preserve a limited subset of historical samples and remain a strong paradigm under bounded memory constraints Rebuffi et al. (2017); Chaudhry et al. (2019); Buzzega et al. (2020). Recent studies further investigate memory construction and sample selection strategies for improving replay efficiency Aljundi et al. (2019); Prabhu et al. (2020). However, existing memory allocation strategies mainly optimize predictive performance, while calibration memory allocation for reliable uncertainty quantification remains unexplored.

Conformal Prediction (CP) constructs prediction sets with distribution-free coverage guarantees under exchangeability Vovk et al. (2005); Shafer & Vovk (2008); Angelopoulos & Bates (2023). For classification tasks, adaptive prediction sets and efficient calibration strategies have been developed for modern neural classifiers Romano et al. (2020); Angelopoulos et al. (2021). Beyond the standard exchangeable setting, weighted and adaptive conformal methods have been proposed to handle covariate shift, temporal distribution shift, and more general non-exchangeability Tibshirani et al. (2019); Gibbs & Candès (2021); Barber et al. (2023).

Recent studies have explored uncertainty calibration in continual learning. Li et al. Li et al. (2024) investigated calibration degradation in continual learning models, while Gao and Liu Gao & Liu (2025) introduced conformal prediction into continual learning for uncertainty quantification. More recently, Feng et al. Feng et al. (2026) studied temporal conformal contracts for uncertainty-tolerant class-incremental learning. However, existing works assume sufficient calibration resources or focus on constructing conformal uncertainty estimates, leaving the calibration memory allocation problem under a fixed budget unexplored.

3 Preliminaries & Problem Setup

CIL setting. We consider a sequence of class-incremental tasks $\{\mathcal{T}_t\}_{t=1}^{T}$, where task $\mathcal{T}_t$ introduces new classes $\mathcal{Y}_t$ and $\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset$ for $i \neq j$. After task t, the observed label space is $\mathcal{Y}_{1:t} = \bigcup_{\tau=1}^{t} \mathcal{Y}_\tau$. Given $\boldsymbol{x}$, the model $f_{\boldsymbol{\theta}_t}$ predicts $p_{\boldsymbol{\theta}_t}(c \mid \boldsymbol{x})$ for $c \in \mathcal{Y}_{1:t}$. Since the model is continually updated, the class-wise predictive distributions and their calibration statistics may also evolve over tasks.

Class-wise conformal calibration. For a labeled example $(\boldsymbol{x}, y)$, we use the nonconformity score

$$s_t(\boldsymbol{x}, y) = 1 - p_{\boldsymbol{\theta}_t}(y \mid \boldsymbol{x}) \in [0, 1]. \tag{1}$$

Let $\mathcal{C}_c^t$ be the m_c^t calibration samples retained for class c, with score set $\mathcal{S}_c^t = \{s_t(\boldsymbol{x}, c) : (\boldsymbol{x}, c) \in \mathcal{C}_c^t\}$. For miscoverage level α, the class-wise threshold is the corrected empirical quantile

$$\widehat{q}_c^t = \mathrm{Quantile}_{\widehat{p}_c^t}(\mathcal{S}_c^t), \qquad \widehat{p}_c^t = \min\left\{1, \frac{\lceil (m_c^t + 1)(1-\alpha) \rceil}{m_c^t}\right\}, \tag{2}$$

and the prediction set is

$$\Gamma_t(\boldsymbol{x}) = \left\{c \in \mathcal{Y}_{1:t} : 1 - p_{\boldsymbol{\theta}_t}(c \mid \boldsymbol{x}) \leq \widehat{q}_c^t\right\}. \tag{3}$$

Problem setup: fixed and irreversible calibration memory. As CIL tasks arrive sequentially, unconstrained retention of calibration samples would make memory consumption grow continuously. Following the fixed-memory setting commonly adopted in replay-based CIL, we impose

$$\sum_{c \in \mathcal{Y}_{1:t}} m_c^t = M_{\mathrm{cal}}, \qquad \mathcal{C}_c^t \subseteq \mathcal{C}_c^{t-1}, \quad m_c^t = |\mathcal{C}_c^t|, \ \forall c \in \mathcal{Y}_{1:t-1}. \tag{4}$$

The second constraint captures irreversibility: once an old-class calibration sample is removed, it cannot be recovered. New classes, in contrast, can construct calibration candidate pools while their current-task data are available. Thus, the problem is to allocate the fixed calibration budget across seen classes while determining how much support to retain under unknown future demand.

4 THEORETICAL ANALYSIS OF CALIBRATION ALLOCATION

Before introducing BACA, we analyze calibration allocation under a fixed and irreversible memory budget. We study three questions: how finite calibration size determines class-wise instability, how a fixed budget should be allocated under heterogeneous demand, and how irreversible deletion affects future reliability. Complete proofs are provided in Appendix B.We fix task t, let $C = |\mathcal{Y}_{1:t}|$, and suppress the task index when no temporal distinction is needed.

Finite-calibration disagreement. For class c, define $S_c(X) = s_t(X, c)$ and let F_c and f_c denote its conditional CDF and density given $Y = c$, with $q_c = F_c^{-1}(1-\alpha)$. Let H_c and h_c denote its CDF and density over the deployment population. Two independent calibration subsets of size m produce thresholds $\hat{q}_c^A$ and $\hat{q}_c^B$, and we define $D_c(m) = \Pr[\mathbb{I}\{S_c(X) \le \hat{q}_c^A\} \neq \mathbb{I}\{S_c(X) \le \hat{q}_c^B\}]$.

Theorem 1 (Finite-calibration disagreement law)**.** *Assume that deployment inputs are independent of the calibration samples, $0 < f_c(q_c) < \infty$, and the regularity conditions for first-moment sample quantile asymptotics in Appendix B.1 hold. Then*

$$D_c(m) = \frac{a_c}{\sqrt{m}} + o(m^{-1/2}), \qquad a_c = 2\sqrt{\frac{\alpha(1-\alpha)}{\pi}}\frac{h_c(q_c)}{f_c(q_c)}. \tag{5}$$

Thus, finite-calibration instability decreases as $m^{-1/2}$, while the class-specific coefficient a_c captures heterogeneous calibration demand. Moreover,

$$\mathbb{E}\big[|\Gamma^A(X) \triangle \Gamma^B(X)|\big] = \sum_{c=1}^{C} D_c(m_c), \qquad J(\mathbf{m}; \mathbf{a}) := \sum_{c=1}^{C} \frac{a_c}{\sqrt{m_c}}, \tag{6}$$

so $J(\mathbf{m}; \mathbf{a})$ characterizes the leading-order prediction-set variation induced by finite calibration sampling.

Oracle allocation under a fixed budget. The preceding result reduces calibration allocation to minimizing $J(\mathbf{m}; \mathbf{a})$ under a fixed total budget.

Theorem 2 (Oracle allocation under a fixed calibration budget)**.** *For $a_c > 0$, $0 < l_c \le u_c$, and $\sum_c l_c \le M \le \sum_c u_c$, the continuous problem $\min_{\mathbf{m}} J(\mathbf{m}; \mathbf{a})$ subject to $\sum_c m_c = M$ and $l_c \le m_c \le u_c$ has the unique global minimizer*

$$m_c^\star = \mathrm{clip}\left[\left(\frac{a_c}{2\lambda}\right)^{2/3}, l_c, u_c\right], \qquad \sum_c m_c^\star = M. \tag{7}$$

When no bound is active,

$$m_c^\star = M\frac{a_c^{2/3}}{\sum_j a_j^{2/3}}, \qquad J^\star(M) = \frac{\left(\sum_c a_c^{2/3}\right)^{3/2}}{\sqrt{M}}. \tag{8}$$

Hence, $m_c^\star \propto a_c^{2/3}$: classes with larger calibration demand should receive more memory. In the interior regime, Equal Allocation is oracle-optimal iff $a_1 = \cdots = a_C$, while $J^\star(M) = \Theta(M^{-1/2})$, revealing diminishing returns from increasing the total calibration budget.

Irreversibility-induced future risk. Theorem 2 assumes a fixed demand state. In CIL, however, the support retained for an old class cannot be increased after deletion. Let $m_c^{(0)} > 0$ denote its support after a current-demand allocation, $m_c^{\mathrm{ref}} > 0$ a reference support, and $\Delta_c^\tau \ge 0$ a future demand increase. Define $\mathcal{E}_c^\tau(m) = \Delta_c^\tau[m^{-1/2} - (m_c^{\mathrm{ref}})^{-1/2}]_+$, where $[z]_+ = \max\{z, 0\}$.

Theorem 3 (Irreversibility-induced future-risk bound)**.** *If $m_c^\tau \le m_c^{(0)}$ for all $\tau > t$, then*

$$\inf_{0 < m \le m_c^{(0)}} \mathcal{E}_c^\tau(m) = \Delta_c^\tau \left[\frac{1}{\sqrt{m_c^{(0)}}} - \frac{1}{\sqrt{m_c^{\mathrm{ref}}}}\right]_+. \tag{9}$$

Thus, compressing a class below its reference support creates a future-demand exposure that cannot be removed by later reallocation. Under the bounded uncertainty model $0 \le \Delta_c^\tau \le \bar{a}^t$, its robust counterpart and the corresponding support-to-demand transformation are

$$R_c^t = \bar{a}^t \left[\frac{1}{\sqrt{m_c^{(0)}}} - \frac{1}{\sqrt{m_c^{\mathrm{ref}}}} \right]_+, \qquad \widetilde{a}_c^t = a_c^t \left(\frac{m_c^{\mathrm{tar}}}{m_c^{(0)}} \right)^{3/2}. \tag{10}$$

The first quantity identifies classes exposed to irreversible future risk, while the second follows from the oracle relation $m_c^\star \propto a_c^{2/3}$. Together, these results motivate estimating current class-wise demand, protecting vulnerable classes before irreversible deletion, and allocating the fixed budget according to the resulting demand.

5 BACA: Budget-Aware Calibration Allocation

Guided by the theoretical results in Sec. 4, we propose **Budget-Aware Calibration Allocation (BACA)** for reliable conformal prediction under fixed and irreversible calibration memory. BACA operationalizes the three principles established above: (1) *Current Demand Estimation* estimates the class-wise sensitivity coefficient characterized by Theorem 1; (2) *Irreversibility-Aware Future Protection* protects classes exposed to the irreversible future risk identified by Theorem 3; and (3) *Marginal Budget Allocation* realizes the resulting protected demand through discrete marginal transfers, following the allocation structure in Theorem 2. Algorithm 1 summarizes BACA, while detailed algorithms are provided in **Appendix A**.

5.1 Current Demand Estimation

Theorem 1 shows that the leading class-wise membership disagreement follows $D_c^t(m) \approx a_c^t/\sqrt{m}$, where a_c^t captures the class-specific calibration sensitivity independently of the retained calibration size. Therefore, Stage 1 estimates a_c^t rather than directly using disagreement observed at a single calibration size.

For each class c, we select K probe sizes $\{m_{c,1}, \ldots, m_{c,K}\}$ within the available calibration range. At each probe size, we draw B independent calibration-subset pairs and compute their class-wise membership disagreement $d_{c,k}^{t,b}$. We then estimate

$$\hat{D}_c^t(m_{c,k}) = \frac{1}{B} \sum_{b=1}^{B} d_{c,k}^{t,b}, \qquad \hat{a}_c^t = \frac{1}{K} \sum_{k=1}^{K} \sqrt{m_{c,k}}\, \hat{D}_c^t(m_{c,k}). \tag{11}$$

The resulting $\hat{a}_c^t$ provides a calibration-size-independent measure of current class-wise demand, with $\hat{D}_c^t(m) = \hat{a}_c^t/\sqrt{m}$ used in the subsequent allocation stages.

5.2 Irreversibility-Aware Future Protection

Theorem 3 shows that compressing a class below a reference support exposes it to future demand growth that cannot be corrected after irreversible deletion. Using $\{\hat{a}_c^t\}$, we first obtain the current-only allocation $\mathbf{m}^{(0)}$ with the procedure in Sec. 5.3. We then define

$$m_c^{\mathrm{ref}} = \begin{cases} m_c^{t-1}, & c \in \mathcal{Y}_{1:t-1}, \\ m_{c,\mathrm{eq}}^t, & c \in \mathcal{Y}_t, \end{cases} \qquad \bar{a}^t = \frac{1}{|\mathcal{Y}_{1:t}|} \sum_{c \in \mathcal{Y}_{1:t}} \hat{a}_c^t. \tag{12}$$

The former preserves the pre-deletion support for old classes, while Equal Allocation provides a neutral reference for new classes. Following Theorem 3, the robust risk is

$$R_c^t = \bar{a}^t \left[\frac{1}{\sqrt{m_c^{(0)}}} - \frac{1}{\sqrt{m_c^{\mathrm{ref}}}} \right]_+. \tag{13}$$

Algorithm 1 Budget-Aware Calibration Allocation (BACA)

Require: Previous calibration memory $\mathcal{M}_{t-1}$, current calibration pools, budget M_{cal}
Ensure: Updated calibration memory $\mathcal{M}_t$
Stage 1: Current Demand Estimation
Estimate class-wise demand coefficients $\{\hat{a}_c^t\}$ from K calibration-size probes
Stage 2: Irreversibility-Aware Future Protection
Obtain current-only allocation $\mathbf{m}^{(0)}$ using $\{\hat{a}_c^t\}$
Compute irreversible risks $\{R_c^t\}$
Select high-risk classes and obtain protected coefficients $\{\tilde{a}_c^t\}$
Stage 3: Allocation under Fixed Budget
Satisfy the minimum calibration quota for newly arriving classes
Iteratively transfer memory from old to new classes according to marginal deletion losses and addition gains
Stop when no beneficial transfer remains and construct $\mathcal{M}_t$
return $\mathcal{M}_t$

We select the highest-risk classes with $R_c^t > 0$, up to $\lceil \rho |\mathcal{Y}_{1:t}| \rceil$ classes. For each selected class, we partially recover its reference support and translate this protection into the demand space:

$$m_c^{\text{tar}} = m_c^{(0)} + \eta [m_c^{\text{ref}} - m_c^{(0)}]_+, \qquad \tilde{a}_c^t = \hat{a}_c^t \left(\frac{m_c^{\text{tar}}}{m_c^{(0)}} \right)^{3/2}. \tag{14}$$

For unprotected classes, $\tilde{a}_c^t = \hat{a}_c^t$. The $3/2$ transformation follows directly from the oracle relation $m_c^\star \propto a_c^{2/3}$ in Theorem 2. Thus, future protection is incorporated into the same demand coefficients used by the final allocation, without introducing an additional hard memory constraint.

5.3 Marginal Budget Allocation

Theorem 2 characterizes the continuous allocation structure induced by the leading-order instability objective. CIL additionally requires integer allocations and irreversible old-class memory. Stage 3 therefore realizes the protected demand $\{\tilde{a}_c^t\}$ through discrete marginal transfers under the constraints in Sec. 3. We use

$$J_t(\mathbf{m}^t) = \sum_{c \in \mathcal{Y}_{1:t}} \frac{\tilde{a}_c^t}{\sqrt{m_c^t}}, \qquad \sum_{c \in \mathcal{Y}_{1:t}} m_c^t = M_{\text{cal}} \tag{15}$$

as the allocation criterion. For one additional or removed calibration sample, the corresponding marginal gain and loss are

$$G_c(m) = \tilde{a}_c^t \left(\frac{1}{\sqrt{m}} - \frac{1}{\sqrt{m+1}} \right), \qquad L_c(m) = \tilde{a}_c^t \left(\frac{1}{\sqrt{m-1}} - \frac{1}{\sqrt{m}} \right). \tag{16}$$

Thus, $G_c(m)$ measures the reduction in modeled instability obtained by adding one sample, whereas $L_c(m)$ measures the increase caused by deleting one.

BACA starts from the actual calibration memory retained after task $t-1$. Each new class first receives the minimum feasible quota $m_{\min}$. If the budget is already full, each required new-class sample is paired with removing one sample from the eligible old class minimizing $L_c(m)$.

After all new classes satisfy $m_{\min}$, BACA performs additional old-to-new transfers. At each step, the receiver and donor are

$$r^\star = \arg \max_{\substack{r \in \mathcal{Y}_t \\ m_r < U_r^t}} G_r(m_r), \qquad d^\star = \arg \min_{\substack{d \in \mathcal{Y}_{1:t-1} \\ m_d > m_{\min}}} L_d(m_d), \tag{17}$$

where U_r^t is the available calibration-pool size for new class r. The transfer $d^\star \to r^\star$ is performed only when $G_{r^\star}(m_{r^\star}) > L_{d^\star}(m_{d^\star})$. Otherwise, no remaining one-sample old-to-new transfer decreases the modeled instability, and the allocation terminates.

This procedure preserves the actual irreversible memory state while allocating new calibration support according to the protected class-wise demand.

Table 1: Results under $\alpha = 0.15$ across all datasets and calibration budgets. Results are averaged over all budget-binding tasks. ΔSize denotes the difference in average prediction-set size between BACA and Equal Allocation (BACA − Equal).

Dataset	Budget	Learner	Jaccard (%) ↑		Disagreement (%) ↓		Coverage (%)		Set Size		
			Equal	BACA	Equal	BACA	Equal	BACA	Equal	BACA	Δ
CIFAR-10	500	ER	93.45	**93.64**	10.87	**10.30**	85.59	85.43	1.12	1.11	-0.01
		DER	93.92	**93.95**	9.09	**8.83**	85.53	85.30	1.02	1.02	-0.01
		MEMO	93.35	**93.57**	9.93	**9.29**	85.15	84.93	1.02	1.01	-0.01
		TagFex	**95.56**	95.52	4.90	**4.83**	85.28	85.11	0.90	0.89	0.00
	1000	ER	95.16	**95.22**	8.33	**8.03**	85.13	85.27	1.15	1.15	0.00
		DER	95.34	**95.42**	7.06	**6.82**	84.95	85.07	1.03	1.04	0.00
		MEMO	95.14	**95.30**	7.24	**6.83**	84.46	84.62	1.03	1.03	0.00
		TagFex	**96.59**	96.59	3.70	**3.66**	85.13	85.12	0.90	0.90	0.00
	2000	ER	96.46	**96.58**	6.48	**6.10**	84.75	84.80	1.21	1.21	0.00
		DER	96.84	**97.02**	5.13	**4.68**	85.05	85.16	1.08	1.08	0.00
		MEMO	96.73	**96.90**	5.11	**4.66**	84.01	84.09	1.07	1.07	0.00
		TagFex	97.88	**97.91**	2.36	**2.30**	84.58	84.64	0.91	0.91	0.00
CIFAR-100	2000	ER	77.34	**81.26**	71.72	**64.65**	86.26	86.71	9.29	9.35	+0.06
		DER	78.54	**83.80**	59.13	**49.86**	85.97	86.56	6.10	6.15	+0.05
		MEMO	76.53	**83.00**	66.05	**55.16**	86.17	86.78	7.07	7.17	+0.10
		TagFex	84.81	**88.06**	35.61	**29.32**	85.38	86.02	2.43	2.43	0.00
	3000	ER	83.42	**88.41**	69.07	**56.78**	86.39	86.48	10.65	10.47	-0.18
		DER	83.91	**91.44**	54.69	**37.39**	86.04	86.07	6.45	6.24	-0.21
		MEMO	83.36	**91.35**	60.26	**41.90**	86.13	86.18	7.58	7.33	-0.24
		TagFex	88.97	**94.10**	30.49	**17.98**	85.47	85.68	2.51	2.44	-0.07
	4000	ER	89.00	**93.96**	61.57	**43.43**	86.03	86.09	11.60	11.59	0.00
		DER	88.53	**96.79**	49.96	**21.20**	85.95	86.17	6.91	6.92	+0.01
		MEMO	88.08	**96.41**	55.66	**26.49**	85.71	85.88	8.11	8.10	-0.01
		TagFex	92.25	**97.37**	25.38	**8.80**	85.14	85.36	2.69	2.68	-0.01
Tiny-ImageNet	4000	ER	77.05	**77.79**	93.48	**93.45**	87.18	87.53	63.93	64.65	+0.72
		DER	74.21	**77.30**	90.85	**87.88**	86.69	87.07	29.47	29.75	+0.27
		MEMO	75.87	**77.98**	93.88	**91.95**	86.58	86.93	36.91	37.14	+0.23
		TagFex	73.87	**77.19**	90.25	**86.94**	86.59	87.12	24.55	24.82	+0.27
	6000	ER	83.24	**84.52**	93.48	**92.55**	87.25	87.05	74.74	73.81	-0.93
		DER	81.31	**85.73**	89.78	**83.76**	86.76	86.71	31.80	31.33	-0.47
		MEMO	82.48	**86.22**	93.70	**89.55**	86.68	86.64	40.45	39.93	-0.52
		TagFex	80.44	**86.26**	90.20	**81.84**	86.94	86.94	26.23	25.68	-0.56
	8000	ER	88.36	**90.54**	93.49	**92.38**	86.65	86.56	86.97	86.70	-0.27
		DER	86.92	**92.93**	88.17	**76.35**	86.24	86.32	33.97	34.02	+0.06
		MEMO	87.75	**92.40**	94.04	**86.06**	86.40	86.53	44.30	44.41	+0.11
		TagFex	86.61	**93.40**	88.30	**73.20**	86.69	86.80	27.99	28.02	+0.03

6 EXPERIMENTS

We evaluate BACA through overall performance evaluation and analyses of the proposed current demand estimation and future protection strategies. Additional ablation and sensitivity studies are provided in the **Appendix C**.

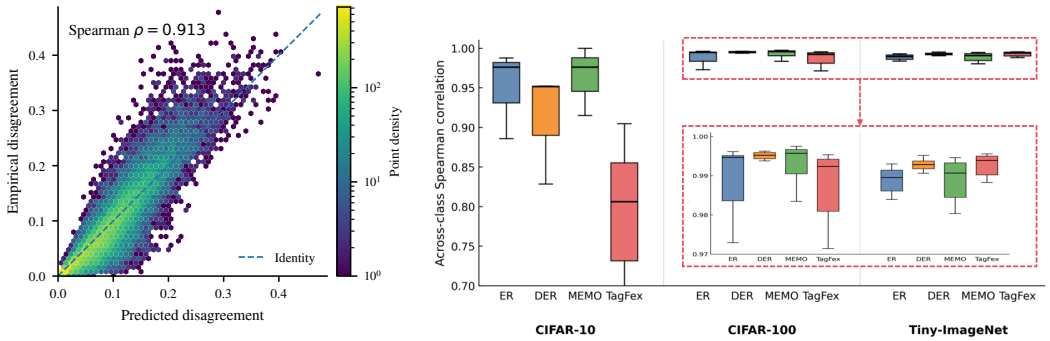
6.1 EXPERIMENTAL SETUP

Datasets and CIL learners. We evaluate BACA on CIFAR-10, CIFAR-100, and Tiny-ImageNet using four representative CIL learners: ER, DER Yan et al. (2021), MEMO Zhou et al. (2023b), and TagFex Zheng et al. (2025), implemented with the PyCIL toolbox Zhou et al. (2023a).

Baseline. For each dataset, we consider three fixed calibration budgets and multiple miscoverage levels α. We compare BACA with Equal Allocation, which uniformly distributes the calibration budget across all seen classes under the same irreversible-memory constraint.

Evaluation metrics. We measure prediction-set stability using Disagreement and Jaccard similarity, and report empirical coverage and average prediction-set size.

Detailed settings and metric definitions are provided in **Appendix C.1**.



(a) Demand estimation (b) Across-class ranking consistency

Figure 2: Validation of current demand estimation. (a) Estimated versus empirical class-wise disagreement. (b) Task-level across-class Spearman correlation between estimated and empirical demand coefficients across datasets and CIL learners.

6.2 Overall Performance under Fixed Calibration Budgets

We evaluate BACA on three benchmarks with four CIL learners under different calibration budgets. Table 1 reports detailed results under the default setting $\alpha = 0.15$, while Table 2 summarizes the overall performance across different miscoverage levels.

Table 2: Overall results across different miscoverage levels, averaged over all datasets, calibration budgets, CIL learners, and budget-binding tasks.

α	Jacc(pp)	Dis (pp)	ΔCov. (pp)	ΔSize
0.05	+3.96	+8.14	+0.07	-0.29
0.10	+3.38	+7.03	+0.15	+0.04
0.15	+3.24	+7.06	+0.13	-0.04
0.20	+3.58	+8.81	+0.13	-0.02
Avg.	**+3.54**	**+7.76**	**+0.12**	**-0.08**

Prediction-set stability. BACA consistently improves prediction-set stability under fixed calibration budgets. Under $\alpha = 0.15$, BACA achieves higher Jaccard similarity and lower disagreement across different datasets, calibration budgets, and CIL learners. BACA improves the average Jaccard similarity by 5.93 percentage points on CIFAR-100 and 3.68 percentage points on Tiny-ImageNet. These results show that class-specific calibration allocation effectively reduces the sensitivity of prediction sets to finite calibration subsets.

Coverage and prediction-set efficiency. The stability gains do not come at the cost of degraded coverage or enlarged prediction sets. Under $\alpha = 0.15$, the average coverage changes by only $+0.13$ percentage points, while the average prediction-set size changes by -0.04. Thus, BACA improves prediction-set stability while preserving both coverage and prediction-set efficiency.

Robustness across miscoverage levels. Table 2 shows that BACA remains effective across $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$. BACA improves average Jaccard similarity by 3.24–3.96 percentage points and reduces disagreement by 7.03–8.81 percentage points, while maintaining comparable coverage and prediction-set size. These results demonstrate that the stability improvement is not specific to a particular miscoverage level. Detailed results are provided in **Appendix C.2**.

6.3 Validation of Current Demand Estimation

We validate the Stage 1 approximation $\hat{D}_c^t(m) = \hat{a}_c^t/\sqrt{m}$ using fresh calibration-subset pairs and independent test data. Figure 2(a) compares the estimated disagreement with the empirical disagreement, where the pooled Spearman correlation reaches 0.913 across classes, tasks, datasets, and CIL learners. This verifies that $\hat{a}_c^t$ can effectively characterize the relationship between calibration size and prediction-set stability.

Since calibration allocation mainly depends on identifying the relative demand among classes, Figure 2(b) further evaluates the consistency of demand ranking. The task-level across-class Spearman correlation between estimated and empirical demand coefficients reaches an average of 0.977. The correlations remain consistently high on CIFAR-100 and Tiny-ImageNet, while CIFAR-10 shows

larger variation due to its smaller number of classes. These results confirm that $\hat{a}_c^t$ provides a reliable class-specific demand signal for subsequent calibration allocation.

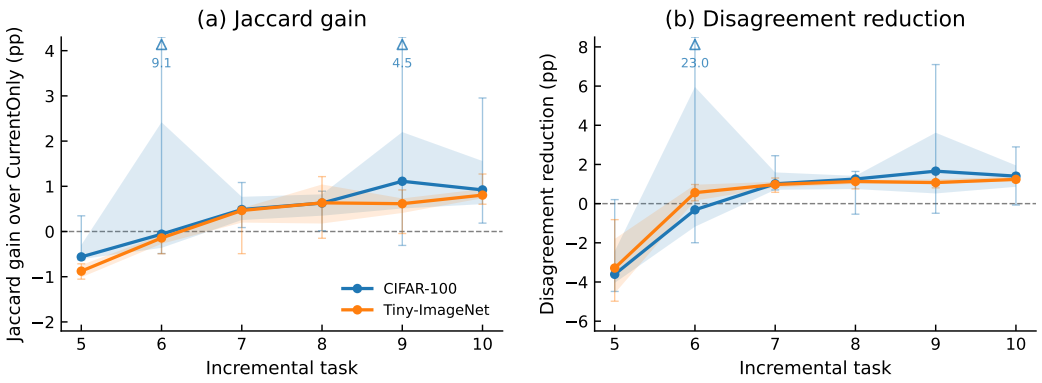


Figure 3: Task-wise contribution of Stage 2 over CurrentOnly across four CIL learners. Shaded regions denote the interquartile range; error bars show min–max values. Positive values indicate improved stability of prediction sets.

6.4 Effect of Future Protection

We isolate the contribution of Stage 2 by comparing full BACA with *CurrentOnly*, which removes future protection while keeping the remaining allocation procedure unchanged. As shown in Table 3, Stage 2 improves both stability metrics at the final task, with an average Jaccard gain of 1.058 percentage points and a 1.346 percentage-point reduction in disagreement. Both metrics improve simultaneously in eight dataset–learner settings, while coverage remains comparable and the average prediction-set size does not increase. Thus, the gain does not rely on larger prediction sets.

Figure 3 further shows that the benefit of Stage 2 mainly emerges in later tasks. In earlier tasks, future protection provides limited or slightly negative short-term gains. As incremental learning proceeds, however, the median changes in both Jaccard similarity and disagreement move toward positive values. This trend is consistent with the role of Stage 2: future protection becomes more relevant as repeated irreversible calibration removal increases the risk of current-demand-only allocation.

Table 3: Final-task contribution of Stage 2 over CurrentOnly. Positive values indicate improved stability.

Dataset	ΔJacc.	ΔDis.	ΔCov.	ΔSize
CIFAR-100	+1.245	+1.408	+0.068	-0.109
Tiny-ImageNet	+0.872	+1.285	-0.256	-1.078
Avg.	**+1.058**	**+1.346**	-0.094	-0.594

7 Conclusion and Future Work

In this paper, we study calibration budget allocation for reliable conformal prediction in class-incremental learning under fixed and irreversible calibration memory. Our analysis characterizes how finite calibration size affects prediction-set instability and how irreversible sample deletion introduces future reliability risk. Based on these insights, we propose BACA, which combines current demand estimation, irreversibility-aware future protection, and marginal budget allocation. Extensive experiments show that BACA consistently improves prediction-set stability while maintaining comparable coverage and prediction-set size across benchmarks, CIL learners, calibration budgets, and miscoverage levels.

Future work includes extending BACA to more dynamic calibration-memory settings where samples may continuously arrive, expire, or become available again. Another direction is to model future calibration demand more adaptively by exploiting longer-term temporal evolution across tasks. It would also be valuable to derive tighter finite-sample characterizations for allocation under very limited calibration budgets. Finally, extending budget-aware calibration allocation to other continual learning settings, such as online or domain-incremental learning, remains an interesting direction.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing, improving clarity and readability, and refining the presentation of the manuscript. We did not use generative AI tools to generate experimental data or to independently conduct experiments or draw scientific conclusions, and other required disclosure tasks are not applicable to this work. All AI-assisted content was carefully reviewed, revised, and verified by the authors to ensure its technical accuracy and consistency with the intended scientific claims. We take full responsibility for the final content of this work, including any text or artifacts produced with the aid of generative AI.

REFERENCES

Rahaf Aljundi, Lucas Caccia, Eugene Belilovsky, Massimo Caccia, Min Lin, Laurent Charlin, and Tinne Tuytelaars. Online continual learning with maximal interfered retrieval. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Anastasios Nikolas Angelopoulos, Stephen Bates, Michael I. Jordan, and Jitendra Malik. Uncertainty sets for image classifiers using conformal prediction. In *International Conference on Learning Representations*, 2021.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276.

Michael Bian and Rina Foygel Barber. Training-conditional coverage for distribution-free predictive inference. *Electronic Journal of Statistics*, 17(2):2044–2066, 2023. doi: 10.1214/23-EJS2145.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: A strong, simple baseline. In *Advances in Neural Information Processing Systems*, volume 33, pp. 15920–15930, 2020.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K. Dokania, Philip H. S. Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

Kun Feng, Xingyu Gao, Xin Qiu, and Rui Guo. Temporal conformal contracts for uncertainty-tolerant class-incremental learning. *SSRN Electronic Journal*, 2026. doi: 10.2139/ssrn.7309608.

Rui Gao and Weiwei Liu. Model uncertainty quantification by conformal prediction in continual learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 18453–18469. PMLR, 2025.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1660–1672, 2021.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. In *Proceedings of the National Academy of Sciences*, volume 114, pp. 3521–3526, 2017.

Klaus-Rudolf Kladny, Bernhard Schölkopf, Lisa Koch, Christian F. Baumgartner, and Michael Muehlebach. A critical perspective on finite sample conformal prediction theory in medical applications. *Artificial Intelligence in Medicine*, 180:103462, 2026. doi: 10.1016/j.artmed.2026.103462.

Lanpei Li, Elia Piccoli, Andrea Cossu, Davide Bacciu, and Vincenzo Lomonaco. Calibration of continual learning models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4160–4169, 2024.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6467–6476, 2017.

Marc Masana, Xialei Liu, Bartlomiej Twardowski, Mikel Menta, Andrew D. Bagdanov, and Joost van de Weijer. Class-incremental learning: Survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2023. doi: 10.1109/TPAMI.2022.3213473.

Marios Papamichalis, Regina Ruane, and Theofanis Papamichalis. Replicable conformal prediction. *arXiv preprint arXiv:2608.23638*, 2026.

Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *European Conference on Computer Vision*, pp. 524–540. Springer, 2020.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Yaniv Romano, Matteo Sesia, and Emmanuel Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3581–3591, 2020.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. *Advances in Neural Information Processing Systems*, 32, 2019.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005. doi: 10.1007/b106715.

Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3014–3023, 2021.

Bowen Zheng, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Task-agnostic guided feature expansion for class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10099–10109, 2025.

Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, and De-Chuan Zhan. Pycil: a python toolbox for class-incremental learning. *Science China Information Sciences*, 66(9):197101, 2023a. doi: 10.1007/s11432-022-3600-y.

Da-Wei Zhou, Qi-Wei Wang, Han-Jia Ye, and De-Chuan Zhan. A model or 603 exemplars: Towards memory-efficient class-incremental learning. In *International Conference on Learning Representations*, 2023b.

Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9851–9873, 2024.

A Detailed Algorithms

This section provides the detailed algorithms for the three stages of BACA. Algorithm 2 estimates the class-wise current calibration demand, Algorithm 3 augments it with irreversibility-aware future protection, and Algorithm 4 realizes the resulting demand through fixed-budget allocation. Algorithm 4 is also used within Stage 2 to obtain the current-only provisional allocation, while the final memory update is performed using the protected demand coefficients.

Algorithm 2 Stage 1: Current Demand Estimation

Require: Available calibration samples $\{\mathcal{A}_c^t\}_{c\in\mathcal{Y}_{1:t}}$, probe sizes $\{m_{c,k}\}_{k=1}^K$, number of repetitions B

Ensure: Current demand coefficients $\{\hat{a}_c^t\}_{c\in\mathcal{Y}_{1:t}}$

for each class $c \in \mathcal{Y}_{1:t}$ **do**
  for $k = 1, \ldots, K$ **do**
    for $b = 1, \ldots, B$ **do**
      Independently sample two calibration subsets $\mathcal{S}_{c,k}^{A,b}$ and $\mathcal{S}_{c,k}^{B,b}$ of size $m_{c,k}$ from $\mathcal{A}_c^t$
      Construct the corresponding conformal prediction sets $\Gamma_t^{A,b}$ and $\Gamma_t^{B,b}$
      Compute class-wise membership disagreement

$$d_{c,k}^{t,b} \leftarrow \Pr_x\Big[\mathbb{I}\{c \in \Gamma_t^{A,b}(x)\} \neq \mathbb{I}\{c \in \Gamma_t^{B,b}(x)\}\Big]$$

    end for

$$\hat{D}_c^t(m_{c,k}) \leftarrow \frac{1}{B}\sum_{b=1}^{B} d_{c,k}^{t,b}$$

  end for

$$\hat{a}_c^t \leftarrow \frac{1}{K}\sum_{k=1}^{K} \sqrt{m_{c,k}}\,\hat{D}_c^t(m_{c,k})$$

end for

return $\{\hat{a}_c^t\}_{c\in\mathcal{Y}_{1:t}}$

Algorithm 3 Stage 2: Irreversibility-Aware Future Protection

Require: Current demand coefficients $\{\hat{a}_c^t\}$, previous calibration memory $\mathcal{M}_{t-1}$, current candidate pools $\{\mathcal{A}_c^t\}$, budget M_{cal}, protection ratio ρ, protection strength η

Ensure: Protected demand coefficients $\{\tilde{a}_c^t\}$

Obtain the current-only allocation

$$\mathbf{m}^{(0)} \leftarrow \textsc{MarginalAllocate}(\{\hat{a}_c^t\}, \mathcal{M}_{t-1}, \{\mathcal{A}_c^t\}, M_{\text{cal}})$$

$\bar{a}^t \leftarrow \dfrac{1}{|\mathcal{Y}_{1:t}|} \displaystyle\sum_{c \in \mathcal{Y}_{1:t}} \hat{a}_c^t$

for each class $c \in \mathcal{Y}_{1:t}$ **do**

  if $c \in \mathcal{Y}_{1:t-1}$ **then**

    $m_c^{\text{ref}} \leftarrow m_c^{t-1}$

  else

    $m_c^{\text{ref}} \leftarrow m_{c,\text{eq}}^t$

  end if

  $R_c^t \leftarrow \bar{a}^t \left[\dfrac{1}{\sqrt{m_c^{(0)}}} - \dfrac{1}{\sqrt{m_c^{\text{ref}}}} \right]_+$

end for

$\mathcal{P} \leftarrow \{c \in \mathcal{Y}_{1:t} : R_c^t > 0\}$

Select the classes with the largest risks:

$$\mathcal{H}_t \leftarrow \text{Top}_{\min(|\mathcal{P}|, \lceil \rho |\mathcal{Y}_{1:t}| \rceil)} \{R_c^t : c \in \mathcal{P}\}$$

for each class $c \in \mathcal{Y}_{1:t}$ **do**

  if $c \in \mathcal{H}_t$ **then**

    $m_c^{\text{tar}} \leftarrow m_c^{(0)} + \eta [m_c^{\text{ref}} - m_c^{(0)}]_+^{3/2}$

    $\tilde{a}_c^t \leftarrow \hat{a}_c^t \left(\dfrac{m_c^{\text{tar}}}{m_c^{(0)}} \right)$

  else

    $\tilde{a}_c^t \leftarrow \hat{a}_c^t$

  end if

end for

return $\{\tilde{a}_c^t\}_{c \in \mathcal{Y}_{1:t}}$

Algorithm 4 Stage 3: Allocation under Fixed Budget

procedure MARGINALALLOCATE($\{a_c\}, \mathcal{M}_{t-1}, \{\mathcal{A}_c^t\}, M_{\text{cal}}$)
  Initialize $m_c \leftarrow |\mathcal{M}_{t-1,c}|$ for $c \in \mathcal{Y}_{1:t-1}$
  Initialize $m_c \leftarrow 0$ for $c \in \mathcal{Y}_t$
  Set $U_c^t \leftarrow |\mathcal{M}_{t-1,c}|$ for old classes and $U_c^t \leftarrow |\mathcal{A}_c^t|$ for new classes

  Mandatory allocation for new classes
  for each new class $r \in \mathcal{Y}_t$ **do**
    while $m_r < m_{\min}$ **do**
      if $\sum_c m_c < M_{\text{cal}}$ **then**
        $m_r \leftarrow m_r + 1$
      else
        $d^\star \leftarrow \arg \min_{\substack{d \in \mathcal{Y}_{1:t-1} \\ m_d > m_{\min}}} a_d \left(\frac{1}{\sqrt{m_d - 1}} - \frac{1}{\sqrt{m_d}} \right)$
        $m_{d^\star} \leftarrow m_{d^\star} - 1$
        $m_r \leftarrow m_r + 1$
      end if
    end while
  end for

  Marginal old-to-new allocation
  while true **do**
    $r^\star \leftarrow \arg \max_{\substack{r \in \mathcal{Y}_t \\ m_r < U_r^t}} a_r \left(\frac{1}{\sqrt{m_r}} - \frac{1}{\sqrt{m_r + 1}} \right)$
    $d^\star \leftarrow \arg \min_{\substack{d \in \mathcal{Y}_{1:t-1} \\ m_d > m_{\min}}} a_d \left(\frac{1}{\sqrt{m_d - 1}} - \frac{1}{\sqrt{m_d}} \right)$
    $G^\star \leftarrow a_{r^\star} \left(\frac{1}{\sqrt{m_{r^\star}}} - \frac{1}{\sqrt{m_{r^\star} + 1}} \right)$
    $L^\star \leftarrow a_{d^\star} \left(\frac{1}{\sqrt{m_{d^\star} - 1}} - \frac{1}{\sqrt{m_{d^\star}}} \right)$
    if $G^\star \leq L^\star$ **then**
      break
    end if
    $m_{r^\star} \leftarrow m_{r^\star} + 1$
    $m_{d^\star} \leftarrow m_{d^\star} - 1$
  end while

  Construct calibration memory
  for each old class $c \in \mathcal{Y}_{1:t-1}$ **do**
    Retain m_c samples from $\mathcal{M}_{t-1,c}$
  end for
  for each new class $c \in \mathcal{Y}_t$ **do**
    Select m_c samples from $\mathcal{A}_c^t$
  end for
  return $\mathbf{m}$ and the resulting calibration memory $\mathcal{M}_t$
end procedure

B Proofs and Additional Theoretical Results

This section provides complete proofs and additional results for the theoretical analysis in Sec. 4. We first prove the finite-calibration disagreement law, then derive the fixed-budget oracle allocation and its main consequences. We next formalize the effect of irreversible calibration deletion under future demand growth. Finally, we analyze robustness to errors in the estimated demand coefficients.

B.1 Finite-Calibration Disagreement

For class c, let Z_c denote the calibration nonconformity score with CDF F_c and density f_c. Let $p = 1 - \alpha$ and $q_c = F_c^{-1}(p)$. The deployment-score CDF for candidate class c is

$$H_c(z) = \Pr_X(S_c(X) \le z)\,, \tag{18}$$

with density h_c.

We assume that q_c is unique, F_c is continuously differentiable in a neighborhood of q_c, and $0 < f_c(q_c) < \infty$. We further assume that H_c is continuously differentiable around q_c with locally Lipschitz derivative, and that deployment inputs are independent of the calibration samples. Finally, the empirical p-quantile is assumed to satisfy the L_1 Bahadur representation

$$\hat{q}_c - q_c = \frac{p - \hat{F}_m(q_c)}{f_c(q_c)} + r_m, \qquad \sqrt{m}\,\mathbb{E}|r_m| \to 0, \tag{19}$$

together with $\sup_m m\,\mathbb{E}[(\hat{q}_c - q_c)^2] < \infty$. These standard moment conditions guarantee the first-moment convergence used below.

Exact disagreement identity. Let $\hat{q}_c^A$ and $\hat{q}_c^B$ be obtained from two independent calibration subsets of size m, and define $I_c^A(X) = \mathbb{I}\{S_c(X) \le \hat{q}_c^A\}$ and $I_c^B(X) = \mathbb{I}\{S_c(X) \le \hat{q}_c^B\}$. Conditioning on the thresholds gives

$$\Pr_X\big[I_c^A(X) \ne I_c^B(X) \mid \hat{q}_c^A, \hat{q}_c^B\big] = \big|H_c(\hat{q}_c^A) - H_c(\hat{q}_c^B)\big|\,. \tag{20}$$

Hence,

$$D_c(m) = \mathbb{E}\,\big|H_c(\hat{q}_c^A) - H_c(\hat{q}_c^B)\big|\,. \tag{21}$$

Asymptotic quantile fluctuation. By Eq. (19) and the central limit theorem,

$$\sqrt{m}\,(\hat{q}_c - q_c) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{f_c(q_c)^2}\right). \tag{22}$$

Since the two calibration subsets are independent,

$$\sqrt{m}\,(\hat{q}_c^A - \hat{q}_c^B) \xrightarrow{d} \mathcal{N}\left(0, \frac{2p(1-p)}{f_c(q_c)^2}\right). \tag{23}$$

The bounded second-moment condition above implies uniform integrability of the normalized quantile differences. Therefore, convergence in distribution in Eq. (23) also yields convergence of the first absolute moment.

Since h_c is locally Lipschitz around q_c, a first-order expansion gives

$$H_c(\hat{q}_c^A) - H_c(\hat{q}_c^B) = h_c(q_c)(\hat{q}_c^A - \hat{q}_c^B) + R_m, \tag{24}$$

where $\mathbb{E}|R_m| = O(m^{-1}) = o(m^{-1/2})$. Combining this expansion with Eq. (21),

$$\begin{aligned}\sqrt{m}\,D_c(m) &\to h_c(q_c)\mathbb{E}\left|\mathcal{N}\left(0, \frac{2p(1-p)}{f_c(q_c)^2}\right)\right| \\ &= \frac{2h_c(q_c)}{f_c(q_c)}\sqrt{\frac{p(1-p)}{\pi}}.\end{aligned} \tag{25}$$

Since $p = 1 - \alpha$, defining

$$a_c = 2\sqrt{\frac{\alpha(1-\alpha)}{\pi}}\frac{h_c(q_c)}{f_c(q_c)} \tag{26}$$

gives

$$D_c(m) = \frac{a_c}{\sqrt{m}} + o(m^{-1/2}), \tag{27}$$

which proves Theorem 1.

Finite-sample conformal quantile. Split conformal prediction uses the corrected order statistic $k_m = \lceil (m+1)(1-\alpha) \rceil$. For feasible m with $k_m \leq m$, its corresponding empirical level $p_m = k_m/m$ satisfies $p_m - p = O(m^{-1})$. Since $f_c(q_c) > 0$, this changes the corresponding population quantile by only $O(m^{-1})$, which is smaller than the $O_p(m^{-1/2})$ sampling fluctuation. Therefore, the conformal quantile correction does not change the leading term in Theorem 1.

From class-wise disagreement to prediction-set variation. For any deployment input X,

$$|\Gamma^A(X) \triangle \Gamma^B(X)| = \sum_{c=1}^{C} \mathbb{I}\left[I_c^A(X) \neq I_c^B(X)\right]. \tag{28}$$

Taking expectation gives the exact identity

$$\mathbb{E}|\Gamma^A(X) \triangle \Gamma^B(X)| = \sum_{c=1}^{C} D_c(m_c). \tag{29}$$

No independence between candidate-class membership indicators is required. For fixed C and $m_c \to \infty$ for all classes,

$$\mathbb{E}|\Gamma^A(X) \triangle \Gamma^B(X)| = \sum_{c=1}^{C} \frac{a_c}{\sqrt{m_c}} + \sum_{c=1}^{C} o(m_c^{-1/2}). \tag{30}$$

This motivates the leading-order objective $J(\mathbf{m}; \mathbf{a}) = \sum_c a_c/\sqrt{m_c}$ used below.

B.2 Oracle Allocation under a Fixed Budget

Consider

$$\min_{\mathbf{m}} J(\mathbf{m}; \mathbf{a}) = \sum_{c=1}^{C} \frac{a_c}{\sqrt{m_c}} \quad \text{s.t.} \quad \sum_{c=1}^{C} m_c = M, \qquad l_c \leq m_c \leq u_c, \tag{31}$$

where $a_c > 0$, $0 < l_c \leq u_c$, and $\sum_c l_c \leq M \leq \sum_c u_c$.

Convexity and KKT conditions. For each class,

$$\frac{\partial J}{\partial m_c} = -\frac{a_c}{2m_c^{3/2}}, \qquad \frac{\partial^2 J}{\partial m_c^2} = \frac{3a_c}{4m_c^{5/2}} > 0. \tag{32}$$

Thus, J is strictly convex over the positive orthant. The feasible region is convex and compact, so the problem has a unique global minimizer.

The Lagrangian is

$$\mathcal{L} = \sum_{c=1}^{C} \frac{a_c}{\sqrt{m_c}} + \lambda\left(\sum_{c=1}^{C} m_c - M\right) + \sum_{c=1}^{C} \mu_c(l_c - m_c) + \sum_{c=1}^{C} \nu_c(m_c - u_c), \tag{33}$$

where $\mu_c, \nu_c \geq 0$. The stationarity condition is

$$-\frac{a_c}{2m_c^{3/2}} + \lambda - \mu_c + \nu_c = 0. \tag{34}$$

For an interior class, $\mu_c = \nu_c = 0$, yielding $m_c = (a_c/(2\lambda))^{2/3}$. Complementary slackness therefore gives

$$m_c^\star = \text{clip}\left[\left(\frac{a_c}{2\lambda}\right)^{2/3}, l_c, u_c\right], \qquad \sum_c m_c^\star = M, \tag{35}$$

for some $\lambda > 0$. The allocation $\mathbf{m}^\star$ is unique, although λ need not be unique when the budget lies exactly on a boundary.

Interior solution. If no bound is active, let $S = \sum_{j=1}^{C} a_j^{2/3}$. The budget constraint gives $(2\lambda)^{-2/3} S = M$, and hence

$$m_c^\star = M \frac{a_c^{2/3}}{S}, \qquad J^\star(M) = \frac{S^{3/2}}{\sqrt{M}}. \tag{36}$$

This proves Theorem 2.

Budget–stability scaling. Within the interior regime,

$$J^\star(\kappa M) = \kappa^{-1/2} J^\star(M), \qquad J^\star(M) = \Theta(M^{-1/2}). \tag{37}$$

Therefore, achieving $J^\star(M) \le \delta$ requires

$$M \ge \frac{\left(\sum_c a_c^{2/3}\right)^3}{\delta^2}. \tag{38}$$

For bounded or irreversible allocation, the feasible set can only be smaller than the unconstrained interior relaxation, so this should be interpreted as an optimistic oracle requirement.

Comparison with Equal Allocation. In the interior regime, $m_c^{\mathrm{eq}} = M/C$ gives

$$J^{\mathrm{eq}} = \sqrt{\frac{C}{M}} \sum_{c=1}^{C} a_c. \tag{39}$$

Define

$$G_{\mathrm{oracle}} = \frac{J^{\mathrm{eq}} - J^\star}{J^{\mathrm{eq}}} = 1 - \frac{\left(\sum_c a_c^{2/3}\right)^{3/2}}{\sqrt{C} \sum_c a_c}. \tag{40}$$

Because $x^{2/3}$ is concave, Jensen's inequality gives

$$\left(\sum_c a_c^{2/3}\right)^{3/2} \le \sqrt{C} \sum_c a_c, \tag{41}$$

with equality if and only if $a_1 = \cdots = a_C$. Consequently,

$$G_{\mathrm{oracle}} \ge 0, \qquad G_{\mathrm{oracle}} = 0 \iff a_1 = \cdots = a_C. \tag{42}$$

Hence, within the interior regime, Equal Allocation is oracle-optimal exactly when all classes have identical calibration sensitivity.

A complementary inequality, $(\sum_c a_c^{2/3})^{3/2} \ge \sum_c a_c$, yields

$$0 \le G_{\mathrm{oracle}} \le 1 - \frac{1}{\sqrt{C}}. \tag{43}$$

For strictly positive a_c, the upper endpoint is approached as the demand becomes increasingly concentrated in a single class.

B.3 Irreversibility-Induced Future Risk

At task t, let a_c^t denote the current demand coefficient and suppose a future task $\tau > t$ has

$$a_c^\tau = a_c^t + \Delta_c^\tau, \qquad \Delta_c^\tau \ge 0. \tag{44}$$

Let $m_c^{(0)} > 0$ be the support retained after the current allocation and $m_c^{\mathrm{ref}} > 0$ a reference support. We isolate the leading-order instability caused by the future demand increase as

$$\mathcal{E}_c^\tau(m; \Delta_c^\tau) = \Delta_c^\tau \left[\frac{1}{\sqrt{m}} - \frac{1}{\sqrt{m_c^{\mathrm{ref}}}}\right]_+. \tag{45}$$

Proof of Theorem 3. Irreversible deletion implies that every future feasible support satisfies $0 < m \le m_c^{(0)}$. Since $m^{-1/2}$ is strictly decreasing,

$$\frac{1}{\sqrt{m}} \ge \frac{1}{\sqrt{m_c^{(0)}}}. \tag{46}$$

Therefore,

$$\mathcal{E}_c^\tau(m;\Delta_c^\tau) \ge \Delta_c^\tau \left[\frac{1}{\sqrt{m_c^{(0)}}} - \frac{1}{\sqrt{m_c^{\mathrm{ref}}}}\right]_+. \tag{47}$$

Equality is attained at $m = m_c^{(0)}$, giving

$$\inf_{0<m\le m_c^{(0)}} \mathcal{E}_c^\tau(m;\Delta_c^\tau) = \Delta_c^\tau \left[\frac{1}{\sqrt{m_c^{(0)}}} - \frac{1}{\sqrt{m_c^{\mathrm{ref}}}}\right]_+. \tag{48}$$

This proves Theorem 3.

Robust characterization under unknown future demand. Suppose the exact future increase is unknown but belongs to

$$0 \le \Delta_c^\tau \le \bar{a}^t, \tag{49}$$

where $\bar{a}^t > 0$ denotes a task-level demand scale. Since $\mathcal{E}_c^\tau$ is nondecreasing in Δ_c^τ and decreasing in m, we obtain

$$\begin{aligned}\inf_{0<m\le m_c^{(0)}} \sup_{0\le\Delta\le\bar{a}^t} \Delta \left[\frac{1}{\sqrt{m}} - \frac{1}{\sqrt{m_c^{\mathrm{ref}}}}\right]_+ \\ = \bar{a}^t \left[\frac{1}{\sqrt{m_c^{(0)}}} - \frac{1}{\sqrt{m_c^{\mathrm{ref}}}}\right]_+.\end{aligned} \tag{50}$$

Thus, the corresponding robust irreversible-risk score is

$$R_c^t = \bar{a}^t \left[\frac{1}{\sqrt{m_c^{(0)}}} - \frac{1}{\sqrt{m_c^{\mathrm{ref}}}}\right]_+. \tag{51}$$

The bounded uncertainty set is a modeling choice used to obtain a robust allocation criterion; it is not implied by conformal coverage theory.

Effect of compression. If $m_c^{(0)} \ge m_c^{\mathrm{ref}}$, then $R_c^t = 0$. Otherwise,

$$\frac{\partial R_c^t}{\partial m_c^{(0)}} = -\frac{\bar{a}^t}{2(m_c^{(0)})^{3/2}} < 0, \qquad \frac{\partial^2 R_c^t}{\partial (m_c^{(0)})^2} = \frac{3\bar{a}^t}{4(m_c^{(0)})^{5/2}} > 0. \tag{52}$$

Hence, the irreversible exposure grows as the retained support decreases, and further deletion becomes increasingly costly when the class is already highly compressed.

Support-to-demand transformation. The interior oracle law in Theorem 2 gives $m_c^\star \propto a_c^{2/3}$. Suppose a protection rule aims to change the unnormalized support weight of class c from $m_c^{(0)}$ toward $m_c^{\mathrm{tar}} > 0$. Let $r_c = m_c^{\mathrm{tar}}/m_c^{(0)}$. Requiring the oracle weight to increase by r_c gives

$$\widetilde{a}_c^{2/3} = r_c a_c^{2/3}, \qquad \widetilde{a}_c = a_c r_c^{3/2} = a_c \left(\frac{m_c^{\mathrm{tar}}}{m_c^{(0)}}\right)^{3/2}. \tag{53}$$

Thus, the exponent $3/2$ follows directly from the $2/3$ oracle allocation law and is not an additional tunable exponent. The transformation applies to the unnormalized oracle weight. After global budget normalization and activation of lower, upper, or irreversible constraints, the final feasible allocation need not equal m_c^{tar} exactly.

B.4 Robustness to Demand-Estimation Error

The preceding analysis uses the true demand coefficients $\mathbf{a}$, whereas BACA estimates them from finite calibration probes. We therefore quantify the effect of multiplicative coefficient error.

[Robustness to demand-estimation error] Let $\mathbf{m}^\star$ minimize $J(\mathbf{m};\mathbf{a})$ over a fixed feasible set $\mathcal{M}$, and let $\widehat{\mathbf{m}}$ minimize $J(\mathbf{m};\widehat{\mathbf{a}})$ over the same set. If

$$(1-\epsilon)a_c \le \widehat{a}_c \le (1+\epsilon)a_c, \qquad 0 \le \epsilon < 1, \tag{54}$$

for every class, then

$$J(\widehat{\mathbf{m}};\mathbf{a}) \le \frac{1+\epsilon}{1-\epsilon} J(\mathbf{m}^\star;\mathbf{a}). \tag{55}$$

Proof. For every feasible allocation $\mathbf{m}$,

$$(1-\epsilon)J(\mathbf{m};\mathbf{a}) \le J(\mathbf{m};\widehat{\mathbf{a}}) \le (1+\epsilon)J(\mathbf{m};\mathbf{a}). \tag{56}$$

By optimality of $\widehat{\mathbf{m}}$ under the estimated coefficients,

$$\begin{aligned}(1-\epsilon)J(\widehat{\mathbf{m}};\mathbf{a}) &\le J(\widehat{\mathbf{m}};\widehat{\mathbf{a}}) \le J(\mathbf{m}^\star;\widehat{\mathbf{a}}) \\ &\le (1+\epsilon)J(\mathbf{m}^\star;\mathbf{a}),\end{aligned} \tag{57}$$

which proves Eq. (55). This argument does not require convexity or continuous allocations and therefore applies to any fixed common feasible set for which the two optimizers are defined.

Sharper interior-regime bound. Let $r_c = \widehat{a}_c/a_c$, so $1-\epsilon \le r_c \le 1+\epsilon$. Under the continuous interior solution,

$$m_c^\star = M\frac{a_c^{2/3}}{\sum_j a_j^{2/3}}, \qquad \widehat{m}_c = M\frac{a_c^{2/3}r_c^{2/3}}{\sum_j a_j^{2/3}r_j^{2/3}}. \tag{58}$$

Therefore,

$$\left(\frac{1-\epsilon}{1+\epsilon}\right)^{2/3} \le \frac{\widehat{m}_c}{m_c^\star} \le \left(\frac{1+\epsilon}{1-\epsilon}\right)^{2/3}. \tag{59}$$

Hence, multiplicative coefficient errors are attenuated through the $2/3$ power in the oracle allocation law.

Let

$$S = \sum_j a_j^{2/3}, \qquad \widehat{S} = \sum_j a_j^{2/3}r_j^{2/3}. \tag{60}$$

Substituting $\widehat{\mathbf{m}}$ into the true objective gives

$$J(\widehat{\mathbf{m}};\mathbf{a}) = \frac{\widehat{S}^{1/2}}{\sqrt{M}}\sum_c a_c^{2/3}r_c^{-1/3}. \tag{61}$$

Since $\widehat{S}^{1/2} \le (1+\epsilon)^{1/3}S^{1/2}$ and $r_c^{-1/3} \le (1-\epsilon)^{-1/3}$,

$$\frac{J(\widehat{\mathbf{m}};\mathbf{a})}{J^\star} \le \left(\frac{1+\epsilon}{1-\epsilon}\right)^{1/3}. \tag{62}$$

This bound is tighter than Eq. (55), but requires the continuous interior regime.

C Additional Experimental Results

C.1 Detailed Experimental Setup

Datasets. We evaluate BACA on three standard class-incremental learning benchmarks. CIFAR-10 is divided into 5 incremental tasks with 2 disjoint classes per task. CIFAR-100 is divided into 10 tasks with 10 classes per task, and Tiny-ImageNet is divided into 10 tasks with 20 classes per task.

CIL learners. All experiments are conducted using the PyCIL toolbox Zhou et al. (2023a). We consider four representative CIL learners: ER, DER, MEMO, and TagFex. For each learner, we follow its corresponding PyCIL implementation and training pipeline.

Calibration settings. We evaluate three fixed calibration budgets for each dataset:

$$\begin{aligned} \text{CIFAR-10: } & M_{\text{cal}} \in \{500, 1000, 2000\}, \\ \text{CIFAR-100: } & M_{\text{cal}} \in \{2000, 3000, 4000\}, \\ \text{Tiny-ImageNet: } & M_{\text{cal}} \in \{4000, 6000, 8000\}. \end{aligned} \tag{63}$$

We further evaluate $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$ to examine robustness across different coverage requirements. Results are averaged over budget-binding tasks, where the fixed calibration budget constrains the number of calibration samples that can be retained.

Baseline. Since existing methods do not directly address calibration-budget allocation for conformal prediction in fixed-budget CIL, we use Equal Allocation as the baseline. At each task, it evenly distributes the available calibration budget among all seen classes and randomly retains calibration samples within each class, subject to the same irreversible-memory constraint as BACA.

Evaluation metrics. Given prediction sets $\Gamma^A(x_i)$ and $\Gamma^B(x_i)$ obtained from two independently sampled calibration subsets, we measure prediction-set stability using Disagreement and Jaccard similarity:

$$\text{Dis.} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\big[\Gamma^A(x_i) \neq \Gamma^B(x_i)\big], \qquad \text{Jacc.} = \frac{1}{N}\sum_{i=1}^{N} \frac{|\Gamma^A(x_i) \cap \Gamma^B(x_i)|}{|\Gamma^A(x_i) \cup \Gamma^B(x_i)|}. \tag{64}$$

Lower Disagreement and higher Jaccard similarity indicate more stable prediction sets across finite calibration subsets.

We additionally report empirical coverage and average prediction-set size:

$$\text{Cov.} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[y_i \in \Gamma(x_i)], \qquad \text{Size} = \frac{1}{N}\sum_{i=1}^{N} |\Gamma(x_i)|. \tag{65}$$

Coverage is expected to remain close to the target level $1 - \alpha$, while average prediction-set size measures prediction-set efficiency.

C.2 MAIN RESULTS OF ADDITIONAL ALPHA

Table 4: Results under $\alpha = 0.05$ across all datasets and calibration budgets. Results are averaged over all budget-binding tasks. ΔSize denotes the difference in average prediction-set size between BACA and Equal Allocation (BACA − Equal).

Dataset	Budget	Learner	Jaccard (%) ↑		Disagreement (%) ↓		Coverage (%)		Set Size		
			Equal	BACA	Equal	BACA	Equal	BACA	Equal	BACA	Δ
CIFAR-10	500	ER	88.52	**89.98**	26.44	**25.09**	95.35	95.72	1.79	1.87	+0.08
		DER	88.50	**89.48**	23.24	**22.55**	95.10	95.39	1.55	1.59	+0.03
		MEMO	88.88	**89.65**	23.72	**22.67**	95.19	95.49	1.56	1.60	+0.04
		TagFex	94.12	**94.62**	11.09	**10.17**	95.45	95.59	1.16	1.15	-0.01
	1000	ER	92.12	**92.26**	20.47	20.47	95.13	95.32	1.84	1.88	+0.03
		DER	92.16	**92.56**	18.39	**17.81**	94.84	94.96	1.56	1.58	+0.01
		MEMO	93.01	**93.28**	16.74	**16.36**	94.87	95.00	1.58	1.60	+0.01
		TagFex	95.64	**96.04**	8.29	**7.59**	95.01	95.24	1.14	1.15	+0.01
	2000	ER	94.14	**94.53**	16.50	**15.68**	94.97	95.04	2.02	2.02	+0.01
		DER	94.38	**95.27**	14.03	**12.27**	94.81	94.88	1.67	1.67	+0.01
		MEMO	95.51	**95.88**	11.55	**10.71**	94.56	94.63	1.69	1.69	0.00
		TagFex	97.04	**97.24**	5.86	**5.37**	94.70	94.75	1.17	1.17	0.00
CIFAR-100	2000	ER	68.65	**74.09**	97.81	**90.86**	96.00	95.87	25.55	24.71	-0.84
		DER	70.13	**77.82**	89.82	**75.38**	95.80	95.61	18.02	17.40	-0.62
		MEMO	69.77	**77.47**	95.09	**82.89**	95.80	95.57	20.76	19.93	-0.83
		TagFex	74.06	**82.03**	68.26	**55.00**	95.43	95.24	8.35	7.92	-0.43
	3000	ER	78.97	**85.53**	97.59	**84.92**	96.75	96.16	33.43	28.17	-5.26
		DER	78.86	**87.19**	86.52	**68.16**	96.55	95.88	22.37	19.39	-2.98
		MEMO	78.94	**87.51**	92.66	**72.26**	96.59	95.95	25.73	22.31	-3.42
		TagFex	79.96	**91.54**	68.90	**40.20**	96.24	95.56	10.87	8.43	-2.44
	4000	ER	85.75	**94.19**	95.25	**70.48**	95.33	95.92	27.34	30.15	+2.81
		DER	88.44	**96.16**	70.07	**45.97**	95.22	95.90	18.32	20.26	+1.94
		MEMO	86.03	**95.97**	80.96	**51.03**	95.22	95.87	21.34	23.67	+2.33
		TagFex	87.05	**97.47**	53.94	**20.80**	94.85	95.78	8.04	9.24	+1.19
Tiny-ImageNet	4000	ER	83.38	**84.43**	99.05	**98.74**	96.40	96.37	105.62	105.45	-0.16
		DER	77.00	**79.89**	97.67	**96.22**	96.11	96.04	68.26	67.57	-0.69
		MEMO	78.98	**81.03**	98.42	**97.49**	96.01	95.97	76.72	76.36	-0.36
		TagFex	75.30	**78.66**	97.95	**96.38**	96.21	96.11	61.82	60.90	-0.92
	6000	ER	89.79	**91.45**	98.23	**93.80**	97.18	96.73	127.82	123.43	-4.39
		DER	84.74	**87.86**	97.55	**93.68**	96.88	96.47	82.45	76.07	-6.39
		MEMO	86.73	**89.08**	98.26	**96.63**	96.81	96.43	92.77	87.16	-5.61
		TagFex	83.68	**88.38**	97.48	**93.59**	96.82	96.44	74.73	67.69	-7.04
	8000	ER	92.73	**95.30**	95.59	**77.99**	95.71	96.36	135.51	137.22	+1.71
		DER	88.45	**94.00**	96.00	**91.80**	95.64	96.44	75.67	77.68	+2.01
		MEMO	89.97	**94.21**	98.48	**95.68**	95.68	96.33	89.14	90.98	+1.85
		TagFex	88.72	**94.81**	95.51	**87.80**	95.70	96.43	66.41	67.43	+1.02

Table 5: Results under $\alpha = 0.10$ across all datasets and calibration budgets. Results are averaged over all budget-binding tasks. ΔSize denotes the difference in average prediction-set size between BACA and Equal Allocation (BACA − Equal).

Dataset	Budget	Learner	Jaccard (%) ↑		Disagreement (%) ↓		Coverage (%)		Set Size		
			Equal	BACA	Equal	BACA	Equal	BACA	Equal	BACA	Δ
CIFAR-10	500	ER	92.35	**92.52**	15.06	**14.75**	90.18	90.34	1.31	1.32	+0.01
		DER	93.19	**93.29**	12.01	**11.94**	89.98	90.25	1.17	1.18	+0.01
		MEMO	92.59	**92.91**	13.30	**12.70**	89.90	90.16	1.18	1.19	+0.01
		TagFex	**95.45**	95.32	**5.95**	6.15	89.94	90.22	0.98	0.98	+0.00
	1000	ER	94.19	**94.39**	12.09	**11.66**	89.95	90.06	1.37	1.38	+0.01
		DER	94.91	**95.19**	9.40	**8.80**	89.77	89.89	1.20	1.20	-0.00
		MEMO	94.25	**94.46**	10.85	**10.30**	89.84	89.92	1.22	1.22	-0.00
		TagFex	**96.61**	96.60	4.52	**4.46**	89.87	89.93	0.99	0.98	-0.00
	2000	ER	96.04	**96.24**	9.08	**8.52**	89.45	89.60	1.46	1.46	+0.01
		DER	96.78	**96.98**	6.42	**5.90**	89.76	89.88	1.26	1.26	+0.00
		MEMO	96.09	**96.21**	7.85	**7.36**	89.33	89.45	1.27	1.28	+0.01
		TagFex	97.74	**97.82**	3.12	**2.96**	89.16	89.30	0.99	1.00	+0.00
CIFAR-100	2000	ER	71.76	**76.57**	85.05	**78.07**	91.01	91.66	13.57	14.32	+0.75
		DER	75.99	**79.74**	69.11	**62.03**	90.76	91.12	9.27	9.72	+0.45
		MEMO	72.97	**79.29**	77.84	**66.72**	91.01	91.34	10.83	11.05	+0.22
		TagFex	79.99	**85.55**	48.23	**38.46**	90.11	90.31	3.75	3.67	-0.08
	3000	ER	81.08	**87.32**	80.55	**66.87**	91.07	91.15	15.07	14.99	-0.08
		DER	82.18	**89.77**	65.41	**47.90**	90.79	90.83	9.90	9.63	-0.27
		MEMO	80.85	**89.70**	72.01	**52.26**	90.92	90.81	11.39	10.97	-0.43
		TagFex	86.60	**93.01**	40.74	**24.70**	90.19	90.07	3.75	3.53	-0.22
	4000	ER	86.47	**93.86**	76.91	**52.25**	90.61	90.65	16.42	16.16	-0.26
		DER	87.67	**95.99**	60.11	**31.04**	90.57	90.47	10.59	10.16	-0.43
		MEMO	87.80	**95.67**	63.68	**36.21**	90.33	90.42	11.93	11.59	-0.34
		TagFex	90.68	**96.98**	34.88	**13.50**	89.82	89.81	3.95	3.75	-0.20
Tiny-ImageNet	4000	ER	78.52	**79.79**	**96.41**	96.45	91.90	92.28	80.58	81.97	+1.39
		DER	73.82	**76.67**	94.27	**92.96**	91.41	91.83	42.34	43.63	+1.28
		MEMO	76.33	**78.01**	96.02	**95.54**	91.24	91.72	50.70	52.50	+1.80
		TagFex	73.17	**76.11**	94.23	**92.62**	91.28	91.92	36.23	38.01	+1.78
	6000	ER	84.55	**85.81**	96.04	**95.81**	91.96	91.95	93.61	93.45	-0.16
		DER	81.08	**85.51**	93.71	**89.06**	91.53	91.44	45.53	44.66	-0.88
		MEMO	83.32	**85.92**	95.98	**93.75**	91.27	91.33	55.23	55.18	-0.05
		TagFex	80.86	**85.84**	93.65	**88.15**	91.60	91.62	38.83	37.98	-0.85
	8000	ER	89.55	**91.76**	95.84	**93.12**	91.31	91.42	107.79	107.65	-0.14
		DER	87.58	**92.66**	91.80	**83.83**	91.02	91.02	47.68	47.30	-0.37
		MEMO	88.32	**92.41**	96.38	**91.40**	91.02	91.09	60.28	59.97	-0.31
		TagFex	86.44	**93.65**	92.89	**80.11**	91.29	91.20	40.87	39.61	-1.25

Table 6: Results under $\alpha = 0.20$ across all datasets and calibration budgets. Results are averaged over all budget-binding tasks. ΔSize denotes the difference in average prediction-set size between BACA and Equal Allocation (BACA − Equal).

Dataset	Budget	Learner	Jaccard (%) ↑		Disagreement (%) ↓		Coverage (%)		Set Size		
			Equal	BACA	Equal	BACA	Equal	BACA	Equal	BACA	Δ
CIFAR-10	500	ER	**93.35**	93.25	**9.36**	9.37	80.46	80.50	0.98	0.98	0.00
		DER	93.80	**94.03**	7.82	**7.42**	80.24	80.17	0.91	0.91	0.00
		MEMO	**94.16**	94.14	7.35	**7.27**	79.66	79.78	0.90	0.90	0.00
		TagFex	**94.81**	94.70	**5.29**	5.38	79.92	80.03	0.83	0.83	0.00
	1000	ER	**94.95**	94.88	7.29	**7.28**	80.10	80.15	1.00	1.00	0.00
		DER	95.42	**95.52**	5.87	**5.60**	79.80	79.86	0.92	0.92	0.00
		MEMO	95.53	**95.65**	5.75	**5.53**	79.27	79.33	0.91	0.92	0.00
		TagFex	**96.10**	96.02	**3.97**	4.04	79.97	80.09	0.83	0.83	0.00
	2000	ER	96.59	**96.74**	5.32	**4.99**	79.90	79.97	1.05	1.05	0.00
		DER	96.95	**97.16**	4.06	**3.72**	80.01	80.03	0.95	0.95	0.00
		MEMO	96.98	**97.00**	4.05	**3.93**	79.00	79.04	0.94	0.94	0.00
		TagFex	97.39	**97.44**	2.67	**2.61**	79.73	79.82	0.84	0.84	0.00
CIFAR-100	2000	ER	79.37	**82.74**	62.88	**56.20**	81.19	81.61	6.68	6.75	+0.07
		DER	80.60	**86.70**	50.51	**38.39**	80.90	81.28	4.11	4.04	-0.07
		MEMO	79.50	**85.11**	55.34	**45.30**	80.92	81.40	4.79	4.81	+0.01
		TagFex	86.36	**89.72**	28.35	**21.70**	80.40	80.74	1.77	1.73	-0.04
	3000	ER	84.35	**89.72**	61.37	**45.77**	81.33	80.97	7.75	7.50	-0.24
		DER	86.09	**93.75**	45.03	**25.17**	81.16	81.02	4.43	4.19	-0.24
		MEMO	85.30	**93.20**	50.23	**29.87**	81.11	80.78	5.24	4.95	-0.28
		TagFex	89.96	**94.74**	23.75	**12.01**	80.48	80.18	1.85	1.77	-0.08
	4000	ER	89.12	**95.69**	54.01	**27.41**	80.23	80.69	8.24	8.33	+0.09
		DER	89.63	**97.65**	39.58	**11.62**	80.19	80.61	4.63	4.59	-0.04
		MEMO	89.38	**96.74**	44.47	**18.72**	79.98	80.49	5.48	5.49	0.00
		TagFex	92.48	**97.45**	19.78	**5.43**	79.33	79.74	1.92	1.92	0.00
Tiny-ImageNet	4000	ER	76.28	**77.61**	90.61	**89.45**	81.94	82.17	50.21	50.62	+0.41
		DER	74.63	**78.44**	87.23	**82.84**	81.43	81.82	20.94	21.05	+0.11
		MEMO	75.72	**78.35**	91.36	**88.63**	81.42	81.72	27.14	27.31	+0.17
		TagFex	74.89	**79.82**	86.61	**79.77**	81.68	82.00	17.19	17.14	-0.05
	6000	ER	82.93	**85.26**	90.31	**87.87**	82.21	81.75	59.43	58.55	-0.87
		DER	81.08	**87.13**	86.47	**75.83**	81.58	81.28	22.69	21.95	-0.75
		MEMO	82.31	**86.92**	91.53	**83.05**	81.62	81.28	30.00	29.25	-0.75
		TagFex	81.67	**88.49**	85.25	**70.98**	82.12	81.86	18.58	17.91	-0.67
	8000	ER	88.98	**91.44**	89.08	**87.36**	81.03	81.58	67.88	69.27	+1.39
		DER	86.88	**94.33**	83.11	**63.17**	80.29	80.87	23.53	23.91	+0.37
		MEMO	87.36	**93.92**	90.27	**73.36**	80.54	80.91	31.81	32.14	+0.33
		TagFex	86.64	**95.09**	83.04	**54.62**	81.26	81.73	19.40	19.63	+0.23

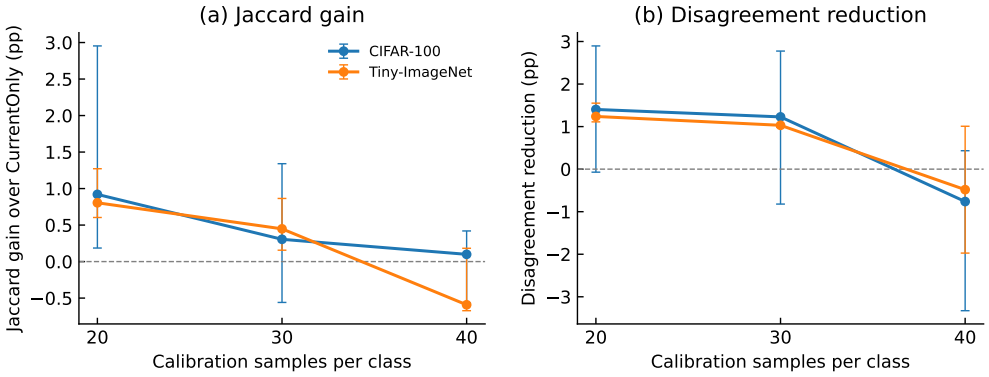


Figure 4: Final-task contribution of Stage 2 over *CurrentOnly* under different calibration budgets. Points denote the median over four CIL learners, and error bars indicate the minimum and maximum learner results. Positive values indicate improved stability.

C.3 Additional Analysis of Future Protection

Effect of the calibration budget. We further examine how the effectiveness of Stage 2 changes with the calibration budget. We fix $\alpha = 0.15$, $\rho = 0.30$, and $\eta = 0.50$, and evaluate three budgets corresponding to 20, 30, and 40 calibration samples per class at the final task. Each setting maintains its complete irreversible allocation trajectory.

Figure 4 shows that future protection is most useful when the calibration budget is tight. With 20 samples per class, Stage 2 improves both Jaccard similarity and prediction-set disagreement on CIFAR-100 and Tiny-ImageNet. Its contribution decreases as the available calibration memory increases and becomes limited or negative under the largest budget. This behavior is consistent with the design of Stage 2. Under a tight budget, more calibration samples must be irreversibly removed, making a current-only allocation more vulnerable to future demand changes. As the budget increases, more calibration support can be retained, reducing the need for additional future protection.

Sensitivity to ρ and η. We next evaluate the sensitivity of the complete BACA framework to the two hyperparameters of Stage 2. We vary the protected-class fraction over $\rho \in \{0.10, 0.20, 0.30, 0.40, 0.50\}$ and the protection strength over $\eta \in \{0.25, 0.50, 0.75\}$. Each configuration maintains its own irreversible allocation trajectory. Results are evaluated at the final task with $\alpha = 0.15$ and averaged over CIFAR-100, Tiny-ImageNet, and four CIL learners.

Figure 5 shows that BACA is insensitive to a broad range of ρ and η. Across all 15 configurations, Jaccard similarity varies only from 70.34% to 70.72%, while prediction-set disagreement varies from 85.42% to 85.91%. Empirical coverage is particularly stable, ranging from 87.53% to 87.66%, and the average prediction-set size varies only from 33.90 to 34.23.

The default setting $(\rho, \eta) = (0.30, 0.50)$ achieves the highest average Jaccard similarity of 70.72%, together with a prediction-set disagreement of 85.52%, coverage of 87.55%, and an average prediction-set size of 33.92. These results indicate that the overall performance is not highly sensitive to the exact choice of ρ and η, while the selected configuration provides a favorable balance among prediction-set stability, coverage, and set size.

C.4 Analysis of Marginal Budget Allocation

We further examine whether Stage 3 translates the protected demand coefficients $\tilde{a}_c^t$ into effective calibration allocations. We consider the final task under $\alpha = 0.15$ on CIFAR-100 and Tiny-ImageNet across all four CIL learners and three calibration budgets.

Figure 6(a) shows a clear relationship between protected demand and the resulting allocation. On CIFAR-100, the median allocation change increases from -44.78% in Q1 to $+41.17\%$ in Q4; on Tiny-ImageNet, it increases from -28.83% to $+24.83\%$. Thus, Stage 3 systematically reallocates

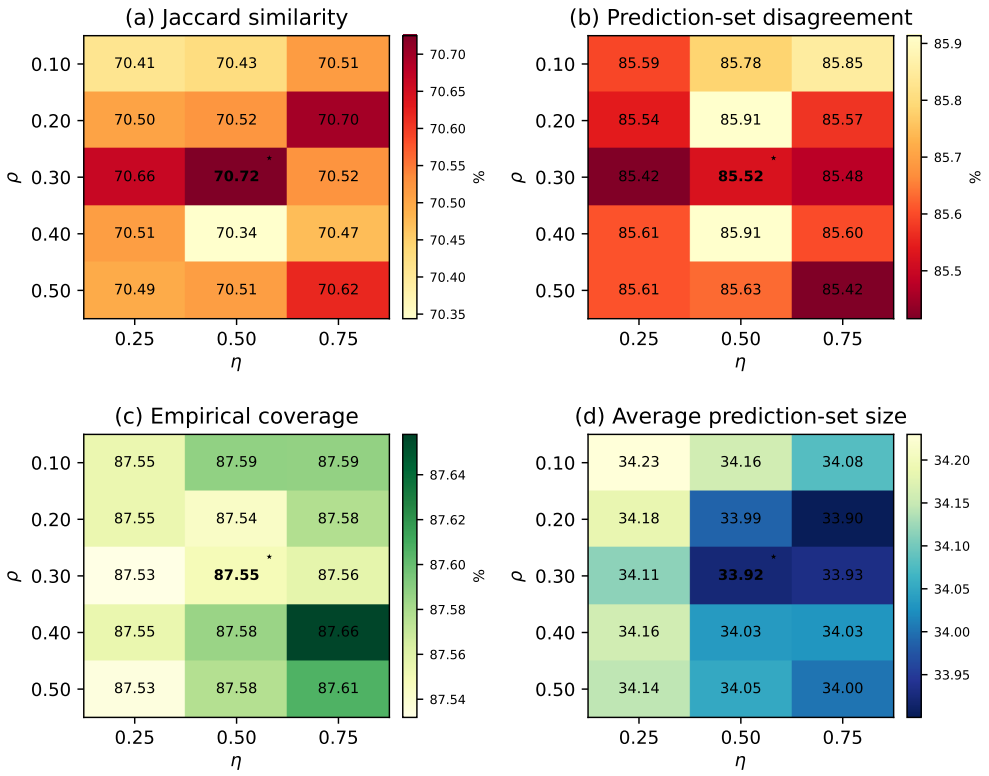


Figure 5: Sensitivity of the final BACA results to ρ and η. Each cell reports the result averaged over CIFAR-100, Tiny-ImageNet, and four CIL learners. The default setting $(\rho, \eta) = (0.30, 0.50)$ is marked with a star. Higher Jaccard similarity is better, while lower prediction-set disagreement and average prediction-set size are preferred at comparable coverage.

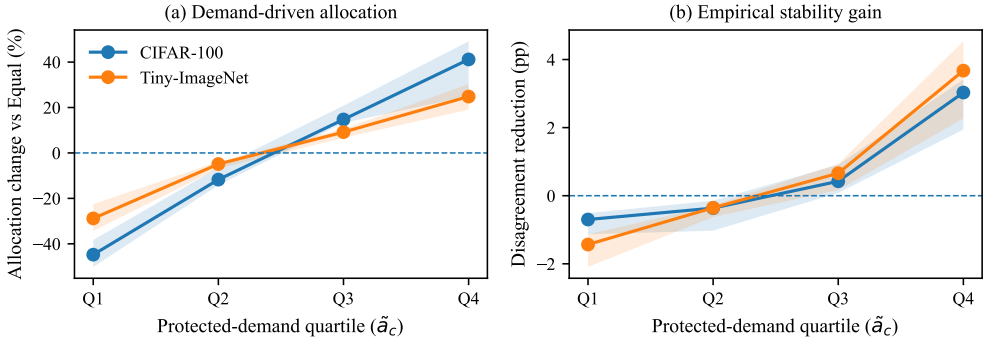


Figure 6: Validation of marginal budget allocation. Classes are grouped into quartiles by their protected demand coefficient $\tilde{a}_c^t$. (a) Relative allocation change from Equal Allocation. (b) Empirical disagreement reduction from Equal Allocation. Curves show the median across learner–budget settings, and shaded regions denote the interquartile range.

calibration samples from lower-demand classes toward higher-demand classes rather than applying a uniform adjustment.

Figure 6(b) shows that this redistribution is also reflected in empirical stability. The median disagreement change increases from -0.70 to $+3.03$ percentage points across Q1–Q4 on CIFAR-100, and from -1.43 to $+3.67$ percentage points on Tiny-ImageNet. Hence, classes receiving more

Table 7: Quantitative validation of Stage 3 at the final task under $\alpha = 0.15$. Spearman correlations are medians across learner–budget settings, while objective reductions are averages. "Positive" denotes settings with a reduced allocation objective.

Dataset	$\rho(\tilde{a}, \Delta m)$	$\rho(\tilde{a}, \Delta D)$	Obj. red. (%)	Positive	Transfers
CIFAR-100	0.943	0.838	5.29	12/12	2,943
Tiny-ImageNet	0.955	0.810	1.94	12/12	5,968
Overall	0.943	0.818	3.62	24/24	8,911

calibration support tend to obtain larger stability improvements, while the fixed budget requires corresponding reductions for lower-demand classes.

Table 7 provides a setting-level quantitative analysis. The median Spearman correlation between protected demand and allocation change reaches 0.943 on CIFAR-100 and 0.955 on Tiny-ImageNet. The corresponding correlations between protected demand and empirical disagreement reduction are 0.838 and 0.810, respectively. These results confirm that the class ordering induced by $\tilde{a}_c^t$ is consistently reflected in both the allocated calibration memory and the resulting stability gain.

We further evaluate the Stage 3 allocation objective in Eq. (15). Compared with Equal Allocation, Stage 3 reduces this objective by an average of 5.29% on CIFAR-100 and 1.94% on Tiny-ImageNet, with reductions in all 24 learner–budget settings. The smallest reduction is still positive at 0.765%.

Finally, the transfer logs provide a direct check of the marginal allocation rule. Across the 24 settings, Stage 3 performs $8{,}911$ accepted marginal transfers. Every accepted transfer has positive net gain, $G_r - L_d > 0$, consistent with the criterion used by Stage 3. Together, these results show that marginal budget allocation converts the protected demand signal into objective-improving transfers and concentrates calibration memory on classes with greater empirical stability benefits.