PSEUDO-LABEL RELIABILITY COLLAPSE: IDENTIFICATION, DECOMPOSITION, AND CALIBRATION FOR CONTINUAL SEMI-SUPERVISED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

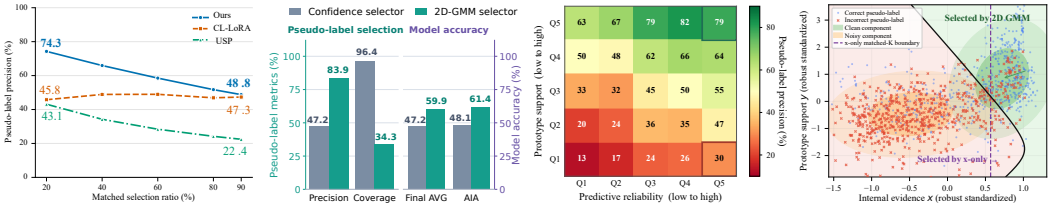
Continual semi-supervised learning (CSSL) faces a critical yet underexplored failure mode beyond catastrophic forgetting, pseudo-label reliability collapse, wherein erroneous pseudo-labels generated under irreversible data access constraints progressively corrupt feature representations across the task stream, precluding the iterative self-correction that underpins standard semi-supervised learning. Existing approaches from both communities are structurally insufficient to address this collapse, as confidence-based pseudo-label selection cannot detect over-confident but semantically inconsistent predictions. To this end, we propose a Reliability-aware Calibration Framework (RCF) grounded in a principled decomposition of pseudo-label reliability into predictive and semantic dimensions. Each unlabeled sample is embedded in a joint predictive-semantic reliability space, where predictive reliability captures augmentation-consistent prediction stability and semantic reliability measures prototype-supported class agreement in the frozen feature space. To estimate the true reliability state of unlabeled samples without ground-truth supervision, we model the resulting reliability representations with a two-component Gaussian Mixture Model whose parameters are inferred via Expectation-Maximization, yielding a calibrated, distribution-relative reliability score that is robust to task-wise distribution shifts. Building on this score, we introduce an adaptive, risk-constrained pseudo-label selection strategy with an asymmetric exponential moving average update rule, which rapidly exploits genuine reliability improvements while conservatively suppressing potential noise propagation. Extensive experiments demonstrate that RCF consistently and significantly outperforms state-of-the-art CSSL methods.

1 INTRODUCTION

Continual learning (CL) enables models to acquire knowledge from sequential non-stationary tasks while preserving previous knowledge Parisi et al. (2019); Ye & Bors (2021). However, most CL methods assume fully labeled data for each task, which is often impractical due to annotation costs. In real-world scenarios, only a small portion of samples are labeled, motivating Continual Semi-Supervised Learning (CSSL) Kang et al. (2023) to address both catastrophic forgetting and label scarcity.

Semi-Supervised Learning (SSL) has demonstrated that unlabeled data can substantially boost model performance when combined with a small labeled set Jiang et al. (2026); Hong et al. (2026). The key mechanism underlying most modern SSL methods is pseudo-labeling that generates class assignments for unlabeled samples based on current model predictions, and then uses these synthetic labels for further optimization Wang et al. (2021). When access to the full training distribution is permitted, repeated refinement cycles allow erroneous pseudo-labels to be corrected over time, producing a virtuous self-improvement loop. CL, on the other hand, studies how models can retain previously acquired knowledge while accommodating newly arriving tasks Zenke et al. (2017). The defining constraint of CL is that historical task data become permanently inaccessible once a new task begins. CSSL sits at the intersection of these two frontiers Kim et al. (2023); Fan et al. (2024). When these two challenges collide, a new, more severe failure mode emerges, which has not been adequately characterized or addressed by either community independently Pan et al. (2020).

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.



(a) Pseudo-Label accuracy (b) Model accuracy (c) heatmap (d) Dimension-compare

Figure 1: Pseudo-label selection on ImageNet-R with 1% labels. (a) Pseudo-label accuracy at different selection ratios, where all methods retain the same number of samples. (b) Comparison between our two-dimensional GMM selector and conventional FixMatch confidence thresholding under the same CL-LoRA framework. (c) Pseudo-label accuracy in the predictive–semantic reliability space.Q1–Q5 denote equal-sized quintiles from low to high. Precision variations within the same predictive group demonstrate the complementary value of semantic evidence. (d) Prediction-only and two-dimensional GMM selection; colors indicate pseudo-label correctness and ellipses denote the two mixture components. More experimental details are provided in Appendix C.

We identify a fundamental failure mode unique to CSSL, termed *pseudo-label reliability collapse*, which refers to the irreversible degradation of pseudo-label quality across tasks. Unlike standard SSL, where pseudo-labels are iteratively refined through repeated access to the full training distribution Kingma et al. (2014); Jiang et al. (2026); Hong et al. (2026); Zhang et al. (2026), CSSL's one-pass data constraint eliminates this correction mechanism Sun et al. (2021). Once a task becomes inaccessible, erroneous pseudo-labels cannot be revised. As a result, inaccurate pseudo-labels degrade feature representations, which in turn reduce the quality of pseudo-labels for subsequent tasks, creating an irreversible feedback loop that accumulates throughout the task stream.

Importantly, this collapse cannot be resolved by confidence-based filtering alone. Deep neural networks are often overconfident on misclassified samples Brahma et al. (2021). Consequently, confidence-based selection, adopted by most existing SSL methods, may retain samples that are semantically inconsistent with the global feature structure, introducing persistent label noise in CSSL Wang et al. (2021). This issue is empirically supported by Fig. 1(a), where representative pseudo-labeling methods achieve limited pseudo-label precision on ImageNet-R, indicating that many selected samples still provide incorrect supervision.

Pseudo-label selection further involves an inherent quantity–quality trade-off. SoftMatch Chen et al. (2023) demonstrates that effective semi-supervised learning depends not only on pseudo-label accuracy, but also on the number of unlabeled samples contributing to optimization. Although both factors are important, we argue that their priorities are asymmetric in CSSL. An incorrectly selected pseudo-label actively introduces a misleading optimization signal, whereas rejecting an unlabeled sample only removes a potential training signal. Since historical samples cannot be revisited, pseudo-label quality should therefore be secured before expanding selection quantity. Fig 1(b) supports this reliability-first principle. Under the same CL-LoRA framework, conventional FixMatch confidence thresholding selects substantially more pseudo-labels Sohn et al. (2020), yet their lower precision introduces noisier supervision and ultimately yields a model with lower final classification accuracy. This comparison indicates that, when pseudo-label quantity and quality cannot be simultaneously maximized, selection precision should be prioritized in CSSL, while the selected set should be expanded only after sufficient reliability has been secured.

Guided by this reliability-first principle, we propose a Reliability-aware Calibration Framework (RCF) for CSSL, aiming to maximize labeling accuracy while simultaneously increasing the number of supervised samples. To achieve this goal, we empirically observe that integrating semantic and predictive characteristics can enhance labelling accuracy, as shown in Fig. 1(c). Based on empirical results, the core of our framework rests on a reliability decomposition hypothesis that the reliability of a pseudo-label can be factored into a predictive reliability component and a semantic reliability component. Predictive reliability captures whether the model's prediction is locally stable and consistent under input perturbations. Semantic reliability captures whether the predicted category is supported by the global semantic structure of the feature space, as encoded in category-level prototypes constructed from labeled data. By representing each unlabeled sample as a point in the resulting predictive-semantic reliability space, our framework transforms pseudo-label selection from a single-score thresholding problem into a latent reliability estimation problem.

To estimate the true reliability distribution from the unlabeled sample pool, we model the reliability space with a two-component Gaussian Mixture Model (GMM), separating the population of unlabeled samples into a reliable component and an unreliable component in a principled, unsupervised manner. As illustrated in Fig. 1(d), incorporating GMM into a reliability space provides a clearer separation between correct and incorrect pseudo-labels that remain overlapped along the prediction-only dimension, demonstrating how the proposed modeling resolves the ambiguity of conventional one-dimensional selection. The posterior probability of belonging to the reliable component, estimated via the Expectation-Maximization algorithm, serves as a calibrated, continuous reliability score for each sample. Crucially, this score is robust to task-wise distribution shifts because it measures the relative position of each sample within the current reliability distribution, rather than comparing against a fixed absolute threshold.

Based on the sample reliability estimation, we introduce an adaptive reliability calibration strategy for pseudo-label selection that automatically adjusts the selection boundary according to a risk constraint on the expected reliability of the selected subset. An asymmetric exponential moving average further stabilizes the selection boundary across training epochs, guided by the asymmetric consequence of reliability errors in which fast exploitation of genuine reliability improvements, but conservative response to potential reliability degradation, which could otherwise trigger irreversible knowledge corruption in the CSSL stream.

To summarize, this paper makes the following contributions: (1) We formally identify and characterize pseudo-label reliability collapse as the fundamental failure mode of CSSL, distinguishing it from catastrophic forgetting and standard SSL noise, and reveal an asymmetric quantity–quality trade-off that motivates prioritizing pseudo-label quality before expanding unlabeled-data utilization. (2) We introduce a principled decomposition of pseudo-label reliability into predictive and semantic dimensions, and propose a predictive-semantic reliability space that enables joint, complementary characterization of pseudo-label quality; (3) We propose a GMM-based latent reliability distribution estimation mechanism that models the reliability state of each unlabeled sample as a latent variable, providing a calibrated, distribution-relative reliability score that is robust to task-wise distribution shifts; (4) We introduce an adaptive, risk-constrained pseudo-label selection strategy with temporal posterior smoothing that dynamically adjusts selection boundaries while conservatively guarding against irreversible pseudo-label noise propagation.

2 Related Work

CSSL learns from sequential tasks with limited labels. Existing methods mitigate forgetting through generative replay and consistency regularization Wang et al. (2021); Brahma et al. (2021), improve unlabeled-data utilization via partitioning and distillation Kang et al. (2023); Fan et al. (2024); Duan et al. (2025), or enhance adaptation with auxiliary networks and LoRA-based isolation Kim et al. (2023); He et al. (2025); Luo et al. (2026). However, pseudo-label reliability is usually treated as an auxiliary factor. Our method explicitly calibrates pseudo-label reliability before unlabeled samples enter continual optimization.

Reliable Pseudo-Label Selection. Reliable pseudo-label selection in SSL commonly starts with confidence thresholding under weak-to-strong consistency, as exemplified by FixMatch Sohn et al. (2020). Subsequent methods replace fixed thresholds with adaptive strategies, including class-wise curricula Zhang et al. (2021), learning-state adaptation Wang et al. (2022), soft sample weighting Chen et al. (2023), and improved class-specific selection Wu & Cui (2024). Beyond confidence, recent approaches exploit semantic similarity or sample-specific reliability estimation Zheng et al. (2022); Liu & Liu (2025; 2026). However, these methods target conventional SSL. In contrast, we model task-dependent reliability in CSSL by jointly leveraging augmentation-stable predictive evidence and frozen-backbone prototype evidence as a task-level latent distribution.

3 Methodology

3.1 Problem Formulation and Network Architecture

Different from conventional CL, where each arriving task usually provides a fully annotated dataset, CSSL considers a more realistic scenario in which only a small proportion of samples are labeled while the majority remain unlabeled. Specifically, the task stream is denoted as $\{\mathcal{T}_1, \ldots, \mathcal{T}_K\}$. For each task $\mathcal{T}_i$, the available data consist of a labeled subset $\mathcal{D}_i = \{(\mathbf{x}_j, y_j)\}_{j=1}^{n_i}$ and an unlabeled subset $\widetilde{\mathcal{D}}_i = \{\mathbf{x}_j\}_{j=1}^{\widetilde{n}_i}$, where n_i and $\widetilde{n}_i$ denote the numbers of labeled and unlabeled samples, respectively. During the sequential learning process, the learner is only allowed to access the current task data $\{\mathcal{D}_i, \widetilde{\mathcal{D}}_i\}$. The observations from previous tasks, i.e., $\{\mathcal{D}_1, \widetilde{\mathcal{D}}_1, \ldots, \mathcal{D}_{i-1}, \widetilde{\mathcal{D}}_{i-1}\}$, are per-

manently unavailable. After optimizing all tasks sequentially, the final model is evaluated on an independent test set to measure the overall continual learning performance.

Although existing CSSL approaches effectively exploit unlabeled samples via pseudo-labeling, they often assume access to the full training distribution. In CSSL, task-wise shifts, limited annotations, and irreversible data constraints can cause early pseudo-label errors to accumulate across tasks, leading to *pseudo-label reliability collapse*, where pseudo-label quality progressively degrades along the task sequence.

Formally, let $R_i(\mathbf{x})$ denote the reliability of pseudo-labels generated at task $\mathcal{T}_i$. In conventional semi-supervised learning, repeated access to training data allows the learner to refine incorrect pseudo-labels. However, in CSSL, the inaccessible historical data prevent such correction:

$$R_{i+1}(\mathbf{x}) \not\vdash R_i(\mathbf{x}) + \Delta R_i(\mathbf{x}), \tag{1}$$

where $\Delta R_i(\mathbf{x})$ denotes the reliability improvement from revisiting previous samples. Inaccurate pseudo-labels may cause representation bias, degrading subsequent pseudo-label quality and leading to cumulative model degradation. Thus, the key challenge in CSSL is not only generating pseudo-labels but also continuously calibrating their reliability in a non-revisitable task stream. Motivated by this, we propose a reliability calibration framework that models pseudo-label reliability from both prediction-level and semantic-level perspectives.

The Network Architecture. Low-Rank Adaptation (LoRA)-based technologies employing a pre-trained Vision Transformer (ViT) as a backbone, have achieved excellent performance in continual learning Hu et al. (2022); He et al. (2025); Luo et al. (2026). In this paper, we adopt a Low-Rank Adaptation (LoRA)-based dynamic expansion architecture He et al. (2025) as the foundation of our approach. Specifically, the proposed architecture utilizes a pre-trained Vision Transformer (ViT) Dosovitskiy et al. (2020) as the backbone network and dynamically introduces a LoRA-based expert module to adaptively learn task-specific knowledge. Let $\mathcal{E}_i$ denote the expert model associated with task $\mathcal{T}_i$. During the training process, only the parameters of the current active expert are updated while all other parameters are frozen.

3.2 Reliability Calibration Framework

Pseudo-label reliability calibration in CSSL is more challenging than in conventional SSL due to the absence of historical data after each task, which prevents correcting erroneous pseudo-labels and causes irreversible representation degradation. Confidence-based selection methods are limited, as prediction confidence only captures local classifier certainty and may conflict with global semantic structures under task shifts. Therefore, reliable pseudo-label estimation requires jointly modeling predictive stability and semantic consistency. Motivated by this, we propose a **Reliability-aware Calibration Framework** to alleviate pseudo-label reliability collapse. The framework is based on the following reliability decomposition hypothesis:

$$R(\mathbf{x}) = [R_{\text{pred}}(\mathbf{x}), R_{\text{sem}}(\mathbf{x})], \tag{2}$$

where R_{pred} measures whether the model prediction is locally trustworthy, while R_{sem} evaluates whether the prediction is supported by the global semantic structure of the feature space. The vector formulation defined in Eq. (2) follows the intuition that pseudo-label reliability requires both predictive stability and semantic agreement. A pseudo-label with either unreliable prediction or insufficient semantic support should be suppressed.

We introduce a **predictive-semantic reliability space**, where pseudo-label reliability for a given sample $\mathbf{x}_j$ is represented from two complementary perspectives:

$$\mathbf{v}_j = (p_j, q_j), \tag{3}$$

where p_j denotes the predictive reliability derived from model behavior, and q_j represents the semantic reliability measured from the feature space. The intuition is that a reliable pseudo-label should satisfy two conditions: (1) the model should consistently predict the same category under perturbations; and (2) the predicted category should be supported by the global semantic structure learned from labeled data. Unlike previous confidence-based approaches that only consider local decision certainty, the proposed reliability space enables explicit identification of over-confident but semantically inconsistent samples.

Robust Category Prototype. To evaluate semantic similarity, we first construct category prototypes from all available labeled samples to capture class-level semantic information. To produce robust

prototypes under limited annotation budgets, each labeled sample is augmented into three weakly-perturbed views $\mathbf{x}_j^{v_1}, \mathbf{x}_j^{v_2}, \mathbf{x}_j^{v_3}$. The corresponding frozen feature representations are then extracted as $\mathbf{f}_j^{v_t} = F_\theta(\mathbf{x}_j^{v_t}),\ t = 1, 2, 3$. Following ℓ_2 feature normalization, the aggregated multi-view representation is computed as $\bar{\mathbf{f}}_j = \mathrm{Normalize}((\hat{\mathbf{f}}_j^{v_1} + \hat{\mathbf{f}}_j^{v_2} + \hat{\mathbf{f}}_j^{v_3})/3)$. For each category c, the prototype is computed as:

$$\tilde{\boldsymbol{\mu}}_c = \mathrm{Normalize}\left(\frac{\sum_{j=1}^{N} F_{\mathrm{eq}}(y_j, c)\bar{\mathbf{f}}_j}{N_c}\right), \tag{4}$$

where $\mathrm{Normalize}(\cdot)$ denotes ℓ_2 feature normalization. $F_{\mathrm{eq}}(\cdot,\cdot)$ returns one when the ground-truth label y_j of a labeled sample equals c, and zero otherwise. N_c denotes the number of labeled samples belonging to category c. Multi-view aggregation reduces augmentation variance and produces a more stable semantic anchor under scarce annotations.

Predictive Reliability Estimation. The first dimension p_j of $\mathbf{v}_j$ evaluates whether the prediction itself is stable and consistent. Given an unlabeled sample $\mathbf{x}_j$,we generate two independent weakly augmented views $\mathbf{x}_j^{w_1}$ and $\mathbf{x}_j^{w_2}$. The expert E_i produces their prediction probability distributions $\mathbf{s}_j^{w_1}$ and $\mathbf{s}_j^{w_2}$ after softmax normalization. We use their average $\bar{\mathbf{s}}_j = (\mathbf{s}_j^{w_1} + \mathbf{s}_j^{w_2})/2$ to determine the pseudo-label $c_j = \arg\max_c \bar{s}_{j,c}$ and its confidence $S_j = \max_c \bar{s}_{j,c}$. However, a single prediction confidence value may be unstable under input perturbations. We therefore measure augmentation inconsistency using the normalized Jensen–Shannon divergence:

$$J_j = \frac{D_{\mathrm{KL}}(\mathbf{s}_j^{w_1} \| \bar{\mathbf{s}}_j) + D_{\mathrm{KL}}(\mathbf{s}_j^{w_2} \| \bar{\mathbf{s}}_j)}{2 \log 2}, \tag{5}$$

where D_{KL} denotes the Kullback–Leibler divergence and $J_j \in [0, 1]$. A smaller value indicates that the prediction remains more stable across the two weak views. The predictive reliability coordinate is then defined as $p_j = \psi(S_j(1 - J_j))$, where $\psi(\cdot)$ denotes the logit transformation $\psi(u) = \log \frac{u}{1-u+\epsilon}$, where ϵ is a small constant. The logit transformation expands the discriminative region near the decision boundary and provides a more informative reliability representation than raw confidence values.

Semantic Reliability Estimation via Prototype Consistency. Prediction confidence only reflects the model's local decision state and does not indicate whether the predicted category agrees with the underlying semantic structure. To address this limitation, we introduce a prototype-based semantic reliability measurement. Different from LoRA-adapted representations, which are continuously modified during task learning, the frozen backbone representation retains stable semantic knowledge from large-scale pretraining. Therefore, it provides a reliable reference space for evaluating whether a pseudo-label is semantically consistent. For an input sample $\mathbf{x}_j$, we reuse the same two weakly augmented views $\mathbf{x}_j^{w_1}$ and $\mathbf{x}_j^{w_2}$ used in predictive reliability estimation. Both views are passed through the complete frozen ViT with all LoRA modules disabled:

$$\mathbf{f}_j^{w_m} = \mathrm{Normalize}\left(F_\theta(\mathbf{x}_j^{w_m})\right), \qquad m \in \{1, 2\}. \tag{6}$$

Their cosine similarities to each class prototype are averaged across the two views:

$$a_{j,c} = \frac{1}{2} \sum\nolimits_{m=1}^{2} \cos\left(\mathbf{f}_j^{w_m}, \widetilde{\boldsymbol{\mu}}_c\right). \tag{7}$$

Given the predicted category c_j, we define semantic agreement as its similarity advantage over the strongest competing category:

$$q_j = a_{j,c_j} - \max_{c \neq c_j} a_{j,c}. \tag{8}$$

A larger positive prototype margin q_j indicates that the frozen semantic space provides stronger support for the predicted category c_j than for any competing category. In contrast, a small or negative q_j indicates that the predicted category has little or no semantic advantage over its strongest competitor, suggesting an ambiguous or potentially incorrect pseudo-label. For notational convenience, let us define $N_u = |\widetilde{\mathcal{D}}_i| = \widetilde{n}_i$ as the number of unique unlabeled samples in the current task, where each original sample is counted only once regardless of repeated sampling during optimization. The coordinate-wise statistics used in the following robust standardization are estimated over these N_u samples, i.e., $\{p_j\}_{j=1}^{N_u}$ and $\{q_j\}_{j=1}^{N_u}$.

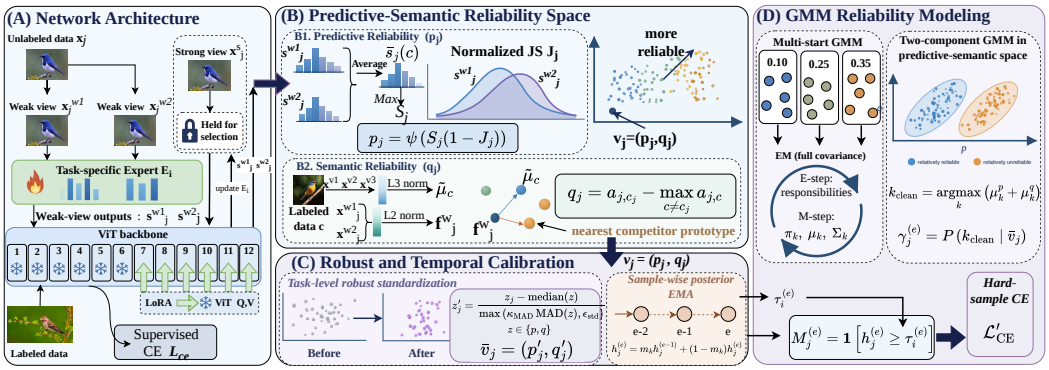


Figure 2: Core framework diagram of the method. **A. Network Architecture**: Two weak views generate reliability evidence and pseudo-labels, while the selected strong view trains the task-specific LoRA expert. **B. Predictive-Semantic Reliability Space**: Predictive stability and frozen-backbone prototype support form a two-dimensional reliability representation. **C. Robust and Temporal Calibration**: Robust coordinate standardization, sample-wise posterior EMA, and epoch-wise threshold smoothing stabilize reliability estimation. **D. GMM Reliability Modeling**: A two-component GMM estimates the reliable posterior, from which selected samples provide hard pseudo-label supervision.

Since the two coordinates have different numerical scales and may contain outliers, we independently normalize them over all unique unlabeled samples of the current task following the modified Z-score formulation Iglewicz & Hoaglin (1993):

$$\begin{aligned} \mathrm{RStd}(p_j) &= \frac{p_j - \mathrm{median}(\mathbf{v}')}{\max(\kappa_{\mathrm{MAD}}\ \mathrm{MAD}(\mathbf{v}'), \beta)}, \qquad \mathbf{v}' \in \{p_1, p_2, \cdots, p_{N_u}\}, \\ \widetilde{\mathbf{v}}_j &= (\mathrm{RStd}(p_j), \mathrm{RStd}(q_j)), \end{aligned} \tag{9}$$

where $\mathrm{median}(\cdot)$ denotes the median and $\mathrm{MAD}(\cdot)$ denotes the median absolute deviation. The constant $\kappa_{\mathrm{MAD}} = 1.4826$ provides Gaussian consistency, and $\beta = 0.05$ is used as a scale floor to avoid unstable normalization under concentrated distributions.Leys et al. (2013) This robust normalization aligns predictive and semantic evidence scales, reducing the impact of outliers and improving GMM stability under task-dependent distribution shifts.

3.3 Latent Reliability Distribution Estimation

After constructing the predictive-semantic reliability space, each unlabeled sample is represented by a reliability vector $\widetilde{\mathbf{v}}_j = (\widetilde{p_j}, \widetilde{q_j}), j = 1, \ldots, N_u$. The proposed reliability space provides a more comprehensive characterization of pseudo-label quality. However, the reliability state of each sample remains unobservable because the true labels of unlabeled data are unavailable. Therefore, pseudo-label reliability is regarded as a latent variable and is estimated through probabilistic distribution modeling. Specifically, we assume that the observed reliability representations are generated from two latent distributions, expressed as:

$$P(\mathbf{v}) = \pi_c P_c(\mathbf{v}) + \pi_n P_n(\mathbf{v}), \tag{10}$$

where P_c and P_n represent the distributions of reliable and unreliable pseudo-labels, respectively. This assumption is motivated by the observation that unlabeled samples in CSSL naturally contain both relatively reliable and noisy pseudo-labels.

Gaussian Mixture Reliability Modeling. Although the predictive-semantic reliability space $\widetilde{\mathbf{v}}_j = (\widetilde{p_j}, \widetilde{q_j})$ provides a richer characterization of pseudo-label quality than scalar confidence, the reliability state of each sample remains fundamentally unobservable. Without access to ground-truth labels, one cannot directly determine whether a pseudo-label is correct. A discriminative approach that directly applies a fixed threshold to a raw reliability score would require a manually specified boundary that is neither adaptive to the current task distribution nor robust to the inter-task distribution shifts inherent in CSSL. More critically, in the absence of ground-truth supervision, such a fixed raw-score boundary implicitly encodes a strong prior about the reliability landscape that may be violated as the task stream evolves.

We therefore adopt a generative perspective in which rather than classifying samples as reliable or unreliable with a fixed rule, we model the population of unlabeled samples as arising from a

mixture of two latent reliability regimes. This framing is motivated by a structural observation of the CSSL setting, where the unlabeled pool at each task inevitably contains reliable and unreliable subsets. The two-component assumption is not merely a modeling convenience. It directly mirrors the binary nature of pseudo-label correctness and yields an interpretable decomposition without over-parametrizing the reliability landscape with unnecessary components. Concretely, we model the joint reliability distribution as a two-component Gaussian Mixture Model (GMM), expressed as:

$$P(\widetilde{\mathbf{v}}_j) = \sum\nolimits_{k=1}^{2} \left\{ \pi_k \mathcal{N}(\widetilde{\mathbf{v}}_j; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}, \tag{11}$$

where π_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ denote the mixture coefficient, mean vector, and covariance matrix of the k-th latent reliability component. The parameters of GMM are optimized through the expectation-maximization (EM) algorithm. During the E-step, the posterior probability that sample $\mathbf{x}_j$ belongs to component k is calculated as:

$$\gamma_{j,k} = \frac{\pi_k \mathcal{N}(\widetilde{\mathbf{v}}_j; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^{2} \pi_l \mathcal{N}(\widetilde{\mathbf{v}}_j; \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}. \tag{12}$$

During the M-step, the effective number of samples assigned to each component is computed as $N_k = \sum_{j=1}^{N_u} \gamma_{j,k}$. The component parameters are then updated as:

$$\pi_k = \frac{N_k}{N_u}, \qquad \boldsymbol{\mu}_k = \frac{\sum_j \gamma_{j,k} \widetilde{\mathbf{v}}_j}{N_k}, \qquad \boldsymbol{\Sigma}_k = \frac{\sum_j \gamma_{j,k} (\widetilde{\mathbf{v}}_j - \boldsymbol{\mu}_k)(\widetilde{\mathbf{v}}_j - \boldsymbol{\mu}_k)^T}{N_k}. \tag{13}$$

Since the number of unlabeled samples in each task may be limited, covariance estimation can become unstable. Therefore, covariance shrinkage is adopted:

$$\boldsymbol{\Sigma}_k \leftarrow (1-\rho)\boldsymbol{\Sigma}_k + \rho \, \mathrm{diag}(\boldsymbol{\Sigma}_k) + \delta \mathbf{I}, \tag{14}$$

where ρ controls shrinkage strength and δ prevents singular covariance estimation. After convergence, the reliable pseudo-label component is identified according to the reliability expectation:

$$k_{\text{clean}} = \arg\max_k \left(\mu_k^p + \mu_k^q \right), \tag{15}$$

where μ_k^p and μ_k^q represent the predictive and semantic reliability centers of component k, respectively. The posterior probability of belonging to the reliable component is then used as the calibrated pseudo-label reliability:

$$\gamma_j^{(e)} = P(k_{\text{clean}}^{(e)} \mid \widetilde{\mathbf{v}}_j^{(e)}). \tag{16}$$

Unlike conventional confidence thresholding, $\gamma_j^{(e)}$ measures the relative position of sample x_j within the reliability distribution fitted at epoch e, rather than comparing its raw confidence with a fixed absolute threshold. All superscripts (e) below indicate quantities computed at the e-th semi-supervised training epoch.

Temporal Posterior Calibration. Since the GMM is re-estimated at each semi-supervised epoch, the posterior reliability of an individual sample may fluctuate with the evolving expert. To obtain a temporally stable reliability estimate, we maintain a sample-wise exponential moving average:

$$h_j^{(e)} = \begin{cases} \gamma_j^{(e)}, & \text{if no historical posterior is available,} \\ m_h h_j^{(e_j^-)} + (1 - m_h)\gamma_j^{(e)}, & \text{otherwise,} \end{cases} \tag{17}$$

where $e_j^- < e$ denotes the most recent epoch in which a valid historical posterior was stored for sample $\mathbf{x}_j$. The coefficient $m_h \in [0, 1)$ controls the posterior momentum, and we set $m_h = 0.7$ in all experiments. When a sample has no historical posterior, its temporal posterior is initialized directly with the current valid GMM posterior $\gamma_j^{(e)}$.

Adaptive Reliability Calibration for Pseudo-label Selection. Although the temporally calibrated posterior $h_j^{(e)}$ provides a stable sample-level reliability estimate, a selection criterion is still required to determine which unlabeled samples participate in consistency optimization. A fixed threshold is unsuitable for CSSL because the optimal reliability level varies across tasks due to annotation scarcity, class imbalance, and domain shift. Therefore, we propose an adaptive calibration strategy

Method	CIFAR-10			CIFAR-100			ImageNet-R			ImageNet-A		
	1%	5%	25%	1%	5%	25%	1%	5%	25%	1%	5%	25%
iCaRL	44.24	44.62	45.02	24.99	26.96	27.78	14.89	20.59	19.50	4.87	15.51	27.06
FOSTER	44.57	44.84	45.07	24.21	26.80	27.67	7.52	11.16	13.46	2.68	6.26	7.20
DER	44.33	44.63	45.02	24.36	26.69	27.50	11.64	14.09	15.22	5.10	9.88	12.12
DER++	44.45	44.68	43.44	20.14	31.82	28.81	5.51	8.60	16.82	8.91	11.14	14.74
DSGD	44.70	45.03	45.09	25.18	27.17	27.96	2.40	2.49	2.22	—	—	—
Keep-LoRA	59.29	59.30	59.11	27.02	28.66	29.62	7.58	9.96	12.28	16.82	26.78	39.36
USP	88.53	90.21	93.07	63.77	70.94	72.51	7.01	27.42	47.67	7.86	13.63	19.03
TACLE	85.78	86.23	86.91	69.98	75.71	77.64	15.02	37.18	45.98	13.39	22.31	33.89
CoVar	94.21	95.26	95.31	**78.33**	79.12	80.11	21.59	52.71	57.14	13.40	25.08	39.18
Ours	**94.97**	**95.35**	**95.39**	76.37	**79.44**	**80.36**	**34.30**	**53.48**	**57.71**	**21.32**	**27.22**	**40.19**

Table 1: Average incremental accuracy (AVG, %) under the same backbone. The best result in each column is highlighted in **bold**.

that determines the selection boundary according to the current reliability distribution. All unlabeled samples are ranked according to their reliability, expressed as:

$$h_{(1)}^{(e)} \geq h_{(2)}^{(e)} \geq \cdots \geq h_{(N_u)}^{(e)}. \tag{18}$$

Instead of manually specifying a confidence threshold, we select the largest subset whose expected reliability satisfies a predefined risk constraint:

$$K^* = \max \left\{ K \,\middle|\, \frac{1}{K} \sum\nolimits_{j=1}^{K} h_{(j)}^{(e)} \geq \eta \right\}, \tag{19}$$

where η denotes the desired reliability level. The candidate threshold at the i-th task learning is therefore obtained as:

$$\tau_i^{\text{candidate}} = h_{(K^*)}^{(e)}. \tag{20}$$

To avoid abrupt threshold fluctuations caused by unstable task observations, we further introduce an asymmetric exponential moving average:

$$\tau_i^{(e)} = \begin{cases} \alpha_u \tau_i^{\text{candidate}} + (1-\alpha_u)\tau_i^{(e-1)}, & \tau_i^{\text{candidate}} > \tau_i^{(e-1)} \\ \alpha_d \tau_i^{\text{candidate}} + (1-\alpha_d)\tau_i^{(e-1)}, & \tau_i^{\text{candidate}} \leq \tau_i^{(e-1)} \end{cases} \tag{21}$$

where $\alpha_u \geq \alpha_d$. The asymmetric update follows the principle that reliability improvement can be quickly exploited, whereas reliability degradation should be treated conservatively because noisy pseudo-labels may cause irreversible knowledge propagation in CSSL. Finally, the selected pseudo-label set is defined as $\mathcal{S}_i = \{\mathbf{x}_j \,|\, h_j^{(e)} \geq \tau_i^{(e)}, j = 1, \cdots, N_u\}$. By using the pseudo-label set $\mathcal{S}_i$ and the labelled dataset $\mathcal{D}_i$, we define an objective function to train the current expert E_i at the i-th task learning:

$$\begin{aligned} \mathcal{L}_{\text{final}} = \mathbb{E}_{(\mathbf{x},y)\sim P_{\mathcal{D}_i}} \big[\sum\nolimits_{s=1}^{C'} \big\{ F_{\text{eq}}(s,y) \log E_i(\mathbf{x})[s] \big\} \big] \\ + \lambda_{\text{sup}} \mathbb{E}_{(\mathbf{x},y)\sim P_{\mathcal{S}_i}} \big[\sum\nolimits_{s=1}^{C'} \big\{ F_{\text{eq}}(s,y) \log E_i(\mathbf{x})[s] \big\} \big], \end{aligned} \tag{22}$$

where $P_{\mathcal{D}_i}$ and $P_{\mathcal{S}_i}$ denote the distribution of $\mathcal{D}_i$ and $\mathcal{S}_i$, respectively. λ_{sup} denotes a hyperparameter that controls the importance of the semi-supervised loss term.

4 Experiments

Baselines and Implementation Details. We compare our method with representative continual-learning and CSSL methods, including iCaRL Rebuffi et al. (2017), Keep-LoRA Luo et al. (2026) , TACLE Kalla et al. (2025), CoVar Liu & Liu (2026) and so on. For a fair comparison, all methods use identical task splits, labeling ratios, and evaluation protocols. Our model employs a pretrained ViT-B/16 with the backbone frozen and inserts task-specific LoRA modules into the query and value projections of the last six Transformer blocks. Further implementation details are provided in **Appendix** C.

4.1 Main Results

Table 1 reports AVG under the same frozen ViT backbone. Our method achieves the best performance in 11 of 12 settings, remaining competitive on CIFAR-10 and leading on CIFAR-100 except under the extremely sparse 1% setting. The largest improvements are observed on ImageNet-R and ImageNet-A, demonstrating the effectiveness of reliability-aware pseudo-label selection under label scarcity and distribution shift.



(a) coordinate-ablation (b) Robust-standardization (c) posterior-precision (d) GMM-Boundary

Figure 3: Visualization of the results of ablation and analysis experiments. More experimental details are provided in Appendix C.

4.2 Ablation Studies

Effectiveness of Predictive–Semantic Reliability Decomposition. As shown in Fig. 3(a), predictive and semantic evidence provide complementary reliability cues across labeling regimes: semantic evidence is particularly informative under scarce supervision, whereas predictive evidence becomes more reliable as supervision increases. Their joint $x + y$ representation therefore forms a unified reliability space, enabling the GMM to adapt its selection boundary to each task distribution.

Effectiveness of Robust Standardization of the Two-Dimensional Reliability Space. We ablate robust standardization by fitting the GMM on raw predictive and semantic coordinates. As shown in Fig. 3(b), removing standardization substantially degrades AVG/AIA under the 1% labeling ratio, while having little effect at 5% and 25%. This indicates that coordinate standardization is primarily beneficial under scarce supervision by improving stable GMM estimation.

Ablation of Posterior EMA and Threshold Smoothing. As shown in Table 2, posterior EMA provides the main improvement by stabilizing sample-level reliability estimates, while threshold smoothing offers complementary boundary refinement. Their combination achieves the best performance under the most challenging setting. The reduced gains with more labeled data indicate that temporal calibration is most beneficial when pseudo-label reliability is highly uncertain.

Table 2: Temporal calibration ablation on ImageNet-R.

EMA	τ Smooth.	1%	5%	25%
No	No	29.23	53.33	57.99
No	Yes	29.47	**53.69**	**58.02**
Yes	No	33.29	53.64	57.80
Yes	Yes	**34.30**	53.48	57.71

4.3 Analytical Experiments

Posterior Reliability versus Raw Confidence. Figure 3 (c) compares pseudo-label ranking precision of S, p, γ, and h on ImageNet-R with 1% labels. At 20% coverage, γ and h achieve 77.1% precision, outperforming S and p (72.6% and 72.8%). The results indicate that GMM posterior improves reliability estimation, while temporal calibration further stabilizes rankings.

Two-Dimensional GMM Partitioning vs. One-Dimensional Confidence Thresholding. We compare a one-dimensional reliability threshold with two-dimensional GMM boundaries. As shown in Fig. 3(d), both variants improve weighted pseudo-label precision and AVG/AIA at 1% labels with similar sample selection. The linear boundary performs slightly better, indicating that semantic evidence provides the main benefit. Gains diminish at higher label ratios, where supervision is stronger. Detailed results are provided in Appendix D.

Table 3: Analysis of prediction stability J on ImageNet-R.

Variant	Metric	1%	5%	25%
Full	AIA	**34.30**	53.48	57.71
	PL ACC.	71.85	94.81	98.56
	Selected	**33.35**	**62.01**	**72.55**
w/o J	AIA	32.59	**53.81**	**57.99**
	PL ACC.	**72.69**	**94.96**	**98.63**
	Selected	31.85	61.11	72.36

Analysis of Augmentation Stability in Predictive Reliability. Table 3 shows that adding augmentation stability J improves AIA from 32.59% to 34.30% and selected fraction from 31.85% to 33.35% under the 1% labeling ratio, while providing limited gains with more labels. This indicates that J is most valuable in low-label settings where confidence estimates are unreliable.

5 Conclusion

We identify pseudo-label reliability collapse in CSSL and propose RCF for reliability-aware pseudo-label calibration. By jointly modeling predictive and semantic reliability, RCF enables robust pseudo-label selection under task shifts and consistently improves performance across multiple benchmarks.

AI USE STATEMENT

Generative AI tools were used for language editing. All technical claims, implementations, experiments, and reported numbers were verified by the authors.

REFERENCES

D. Brahma, V. K. Verma, and P. Rai. Hypernetworks for continual semi-supervised learning. *arXiv preprint arXiv:2110.01856*, 2021.

Hongxin Chen, Runzhe Tao, Yifan Fan, Yidong Wang, Jian Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. *arXiv preprint arXiv:2301.10921*, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yue Duan, Taicai Chen, Lei Qi, and Yinghuan Shi. Divide-and-conquer for enhancing unlabeled learning, stability, and plasticity in semi-supervised continual learning. In *IEEE/CVF International Conference on Computer Vision*, 2025.

Yan Fan, Yu Wang, Pengfei Zhu, Dongyue Chen, and Qinghua Hu. Persistence homology distillation for semi-supervised continual learning. In *Advances of Neural Information Processing Systems*, 2024.

Jiangpeng He, Zhihao Duan, and Fengqing Zhu. Cl-lora: Continual low-rank adaptation for rehearsal-free class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 30534–30544. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02843. URL https://openaccess.thecvf.com/content/CVPR2025/html/He_CL-LoRA_Continual_Low-Rank_Adaptation_for_Rehearsal-Free_Class-Incremental_Learning_CVPR_2025_paper.html.

Yundi Hong, Ao Li, Xinyuan Song, Yan Zhong, Hongyang He, and Victor Sanchez. Beyond data augmentation: Energy-based kuramoto neurons for semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2913–2922, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Boris Iglewicz and David C Hoaglin. *Volume 16: how to detect and handle outliers*. Quality Press, 1993.

Bingbing Jiang, Jie Wen, Zidong Wang, Weiguo Sheng, Zhiwen Yu, Huanhuan Chen, and Weiping Ding. Scalable semi-supervised learning with discriminative label propagation and correction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Jayateja Kalla, Rohit Kumar, and Soma Biswas. Tacle: Task and class-aware exemplar-free semi-supervised class incremental learning. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 6944–6954, February 2025.

Zhiqi Kang, Enrico Fini, Moin Nabi, Elisa Ricci, and Karteek Alahari. A soft nearest-neighbor framework for continual semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11868–11877, October 2023.

Sanghwan Kim, Lorenzo Noci, Antonio Orvieto, and Thomas Hofmann. Achieving a better stability-plasticity trade-off via auxiliary networks in continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *Proc. Advances in Neural Inf. Proc. Systems (NIPS)*, pp. 3581–3589, 2014.

Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology*, 49(4):764–766, 2013.

Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020.

Jinshi Liu and Pan Liu. A confidence-variance theory for pseudo-label selection in semi-supervised learning. *arXiv e-prints*, pp. arXiv–2601, 2026.

Pan Liu and Jinshi Liu. When confidence fails: revisiting pseudo-label selection in semi-supervised semantic segmentation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21874–21884. IEEE, 2025.

Mao-Lin Luo, Zi-Hao Zhou, Yi-Lin Zhang, Yuanyu Wan, Tong Wei, and Min-Ling Zhang. Keeplora: Continual learning with residual gradient adaptation. *CoRR*, abs/2601.19659, 2026. doi: 10.48550/ARXIV.2601.19659. URL https://doi.org/10.48550/arXiv.2601.19659.

Pingbo Pan, Siddharth Swaroop, Alexander Immer, Runa Eschenhagen, Richard Turner, and Mohammad Emtiyaz E Khan. Continual deep learning by functional regularisation of memorable past. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4453–4464, 2020.

G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

Karl Pearson. Iii. contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. (A.)*, (185):71–110, 12 1894. ISSN 0264-3820. doi: 10.1098/rsta.1894.0003. URL https://doi.org/10.1098/rsta.1894.0003.

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

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin D Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, pp. 596–608, 2020.

Gan Sun, Yang Cong, Jiahua Dong, Yuyang Liu, Zhengming Ding, and Haibin Yu. What and how: Generalized lifelong spectral clustering via dual memory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

Abraham Wald. Note on the Consistency of the Maximum Likelihood Estimate. *The Annals of Mathematical Statistics*, 20(4):595 – 601, 1949. doi: 10.1214/aoms/1177729952. URL https://doi.org/10.1214/aoms/1177729952.

Lijie Wang, Kaiyang Yang, Chen Li, Liang Hong, Zhi Li, and Jun Zhu. ORDiSCo: Effective and efficient usage of incremental unlabeled data for semi-supervised continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5383–5392, 2021.

Yidong Wang, Hongxin Chen, Qian Heng, Wei Hou, Yifan Fan, Zuxuan Wu, Jian Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022.

Zhiyu Wu and Jinshi Cui. Allmatch: exploiting all unlabeled data for semi-supervised learning. *arXiv preprint arXiv:2406.15763*, 2024.

Sidney J. Yakowitz and John D. Spragins. On the Identifiability of Finite Mixtures. *The Annals of Mathematical Statistics*, 39(1):209 – 214, 1968. doi: 10.1214/aoms/1177698520. URL https://doi.org/10.1214/aoms/1177698520.

Fei Ye and Adrian G. Bors. Lifelong infinite mixture model based on knowledge-driven Dirichlet process. In *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 10695–10704, 2021.

F. Zenke, B. Poole, and S. Ganguli. Continual learning through synaptic intelligence. In *Proc. of Int. Conf. on Machine Learning, vol. PLMR 70*, pp. 3987–3995, 2017.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in neural information processing systems*, 34:18408–18419, 2021.

Ping Zhang, Zheda Mai, Quang-Huy Percy Nguyen, and Wei-Lun Harry Chao. Revisiting semi-supervised learning in the era of foundation models. *Advances in Neural Information Processing Systems*, 38:59295–59324, 2026.

Mingkai Zheng, Shan You, Lang Huang, Fei Wang, Chen Qian, and Chang Xu. Simmatch: Semi-supervised learning with similarity matching. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14451–14461. IEEE, 2022.

A Additional Discussions on Related Work

A.1 Reliability Evidence Beyond Confidence

Prior work has shown that confidence alone is insufficient for reliable pseudo-label selection. SimMatch combines semantic and instance-level similarities to improve matching targets Zheng et al. (2022). More directly, Liu and Liu construct a confidence-distribution feature space with sample-specific selection boundaries for semi-supervised segmentation Liu & Liu (2025). CoVar combines maximum confidence with residual-class variance for pseudo-label selection Liu & Liu (2026). These studies establish the value of multi-dimensional evidence; we do not claim to be the first to use a two-dimensional space. Our distinction is to combine cross-view prediction stability with frozen-backbone prototype support and re-estimate their joint reliability distribution for each incoming class-incremental task.

A.2 Mixture-Based Sample Partitioning

DivideMix fits a mixture model to per-sample losses to separate potentially clean and noisy labeled examples Li et al. (2020). In contrast, our GMM models the predictive–semantic evidence of unlabeled samples rather than their training losses. Its posterior is further calibrated across epochs before pseudo-label selection. Thus, our contribution lies in the modeled evidence and its use in the continual setting, not in introducing GMM fitting itself.

A.3 Pseudo-Labeling in Continual Semi-Supervised Learning

Existing CSSL methods also address pseudo-label quality. In particular, USP uses divide-and-conquer pseudo-labeling to exploit both high- and low-confidence unlabeled samples, alongside mechanisms for stability and plasticity Duan et al. (2025). Our focus differs: we explicitly estimate sample-wise reliability from complementary predictive and semantic evidence, then calibrate the posterior and selection threshold as the current task model evolves.

A.4 Pseudo-Label Quality, Coverage, and Downstream Learning

SoftMatch emphasizes that effective semi-supervised learning depends on both pseudo-label quality and the amount of unlabeled supervision Chen et al. (2023). In CSSL, incorrect pseudo-labels may introduce biased training signals that cannot be directly corrected by revisiting past-task samples. However, maximizing selection precision alone is also insufficient: overly restrictive selection can leave too few samples to improve the model. We therefore assess pseudo-label precision together with selection coverage and downstream AVG/AIA. Matched-budget comparisons further isolate differences in selection quality from differences in the number of selected samples.

B Theoretical Analysis

In this section, we provide theoretical justifications for the proposed Reliability-aware Calibration Framework. Specifically, we analyze: (i) the irreversibility of pseudo-label reliability degradation in

CSSL (Section B.1); (ii) the identifiability of the two-component GMM in the predictive-semantic reliability space (Section B.2); and (iii) the convergence of the adaptive threshold under the asymmetric exponential moving average (Section B.3).

B.1 Pseudo-label Reliability Collapse in Non-Revisitable Streams

We first formalize the notion of *pseudo-label reliability collapse* and show that it is an inherent consequence of the non-revisitable data constraint in CSSL, rather than a deficiency of any particular algorithm.

Definition B.1 (Task-wise Reliability) *Let $f_i : \mathcal{X} \to \Delta^{C-1}$ denote the classifier induced by the expert $\mathcal{E}_i$ after training on task $\mathcal{T}_i$, where Δ^{C-1} is the $(C-1)$-simplex. The pseudo-label reliability of an unlabeled sample $\mathbf{x}$ at task $\mathcal{T}_i$ is defined as*

$$R_i(\mathbf{x}) \;=\; \mathbb{P}_y\Big(\arg\max_c f_i(\mathbf{x})[c] = y \;\Big|\; \mathbf{x}\Big), \tag{23}$$

where y is the ground-truth label of $\mathbf{x}$.

Definition B.2 (Reliability Improvement Operator) *For a learner with access to a dataset $\mathcal{D}$, define the reliability improvement operator $\Delta R(\,\cdot\,;\mathcal{D})$ as the expected gain in reliability achieved by one additional optimization step using $\mathcal{D}$:*

$$\Delta R_i(\mathbf{x};\mathcal{D}) \;=\; \mathbb{E}\Big[R_{i+1}(\mathbf{x}) - R_i(\mathbf{x}) \;\Big|\; \mathcal{D}\Big]. \tag{24}$$

Proposition B.3 (Reliability Non-recoverability) *Under the non-revisitable data constraint of CSSL, i.e., $\{\mathcal{D}_k, \widetilde{\mathcal{D}}_k\}_{k<i}$ are unavailable when learning task $\mathcal{T}_i$, the reliability improvement operator satisfies*

$$\mathbb{E}\Big[\Delta R_i\Big(\mathbf{x}; \bigcup\nolimits_{k=1}^{i}\{\mathcal{D}_k, \widetilde{\mathcal{D}}_k\}\Big)\Big] \;\geq\; \mathbb{E}\Big[\Delta R_i(\mathbf{x}; \mathcal{D}_i \cup \widetilde{\mathcal{D}}_i)\Big], \tag{25}$$

with strict inequality whenever the historical tasks introduce a non-trivial reliability gradient on $\mathbf{x}$.

Denote $\mathcal{H}_i = \bigcup_{k=1}^{i-1}\{\mathcal{D}_k, \widetilde{\mathcal{D}}_k\}$ and $\mathcal{C}_i = \mathcal{D}_i \cup \widetilde{\mathcal{D}}_i$. The gradient of the cross-entropy loss ℓ with respect to the model parameters θ can be decomposed as

$$\nabla_\theta \mathcal{L}(\theta; \mathcal{H}_i \cup \mathcal{C}_i) = \nabla_\theta \mathcal{L}(\theta; \mathcal{H}_i) + \nabla_\theta \mathcal{L}(\theta; \mathcal{C}_i). \tag{26}$$

Under the standard smoothness assumption that $\mathcal{L}$ is L-smooth in θ, one gradient step with the full data yields a loss reduction of at least

$$\mathcal{L}(\theta; \mathcal{H}_i \cup \mathcal{C}_i) - \mathcal{L}(\theta^+; \mathcal{H}_i \cup \mathcal{C}_i) \;\geq\; \frac{1}{2L}\left\|\nabla_\theta \mathcal{L}(\theta; \mathcal{H}_i \cup \mathcal{C}_i)\right\|^2, \tag{27}$$

while the reduction achievable using $\mathcal{C}_i$ alone satisfies

$$\frac{1}{2L}\left\|\nabla_\theta \mathcal{L}(\theta; \mathcal{C}_i)\right\|^2 \;\leq\; \frac{1}{2L}\left\|\nabla_\theta \mathcal{L}(\theta; \mathcal{H}_i \cup \mathcal{C}_i)\right\|^2, \tag{28}$$

where the last inequality follows from the non-negativity of $\left\|\nabla_\theta \mathcal{L}(\theta; \mathcal{H}_i)\right\|^2$ and the Cauchy–Schwarz bound on the cross term. Since $R_i(\mathbf{x})$ is a monotone function of the loss decrease on a neighborhood of $\mathbf{x}$ (under the standard regularity of the posterior probability), the reliability improvement satisfies the stated inequality. Strict inequality holds whenever $\nabla_\theta \mathcal{L}(\theta; \mathcal{H}_i) \neq \mathbf{0}$, i.e., when historical data carry non-trivial gradient information.

Proposition B.3 shows that the reliability gap introduced at each task is structurally irreversible under the CSSL protocol: the missing historical gradient prevents the learner from recovering the reliability improvement that would otherwise be achievable. This justifies the need for a proactive reliability calibration strategy that compensates for the absent historical correction signal.

B.2 Identifiability and Separation of the Two-Component GMM

We now establish conditions under which the two-component GMM fitted in the predictive-semantic reliability space is identifiable, and show that the reliable component can be consistently recovered.

Assumption B.4 (Reliability Space Separability) *There exist constants $\Delta_\mu > 0$ and $\sigma_{\max} < \infty$ such that the two latent reliability components satisfy*

$$\|\boldsymbol{\mu}_{k_1} - \boldsymbol{\mu}_{k_2}\|_2 \geq \Delta_\mu, \qquad \|\boldsymbol{\Sigma}_k\|_{\text{op}} \leq \sigma_{\max}^2, \quad k = 1, 2,\ k_1 \neq k_2. \tag{29}$$

Theorem B.5 (GMM Identifiability) *Under Assumption B.4 and the standard identifiability conditions for Gaussian mixtures Pearson (1894), the parameter set $\Theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^2$ is identifiable from the marginal distribution $P(\widetilde{\mathbf{v}})$ up to label permutation.*

For a finite mixture of multivariate Gaussians in $\mathbb{R}^d$ with $d \geq 1$, identifiability (up to permutation of components) holds whenever all component means are distinct Yakowitz & Spragins (1968). Assumption B.4 guarantees $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, and the bounded operator norms ensure that each Gaussian component is non-degenerate. Therefore, the classical identifiability result applies, and Θ is uniquely determined by $P(\widetilde{\mathbf{v}})$ up to label permutation.

Corollary B.6 (Consistent Recovery of the Reliable Component) *Under the conditions of Theorem B.5, let the reliable component be defined as $k_{\text{clean}} = \arg\max_k(\mu_k^p + \mu_k^q)$. Given N_u i.i.d. samples from $P(\widetilde{\mathbf{v}})$, the maximum-likelihood estimator $\hat{\Theta}_{N_u}$ satisfies $\hat{\Theta}_{N_u} \to \Theta$ almost surely as $N_u \to \infty$ (by the consistency of MLE for identifiable parametric families Wald (1949)). Consequently, the estimated reliable component $\hat{k}_{\text{clean}}$ coincides with k_{clean} for all sufficiently large N_u.*

The covariance shrinkage step $\boldsymbol{\Sigma}_k \leftarrow (1-\rho)\boldsymbol{\Sigma}_k + \rho\,\text{diag}(\boldsymbol{\Sigma}_k) + \delta\mathbf{I}$ enforces a strictly positive definite lower bound on the smallest eigenvalue of each component covariance, thereby satisfying the non-degeneracy requirement in Theorem B.5 even when the number of unlabeled samples N_u is small relative to the feature dimension.

B.3 Convergence of the Adaptive Reliability Threshold

We analyze the convergence behavior of the asymmetric exponential moving average (EMA) threshold and show that it stabilizes around the oracle threshold induced by the stationary task distribution.

Definition B.7 (Oracle Threshold) *Given the desired reliability level η and the stationary task distribution $P_i^*(\widetilde{\mathbf{v}})$, the oracle threshold at task $\mathcal{T}_i$ is defined as*

$$\tau_i^* = \inf\left\{\tau \;\middle|\; \mathbb{E}_{P_i^*}\Big[h(\widetilde{\mathbf{v}}) \;\Big|\; h(\widetilde{\mathbf{v}}) \geq \tau\Big] \geq \eta\right\}, \tag{30}$$

where $h(\widetilde{\mathbf{v}}) = P(k_{\text{clean}} \,|\, \widetilde{\mathbf{v}})$ denotes the oracle posterior reliability.

Theorem B.8 (Threshold Convergence) *Assume that the candidate threshold sequence $\{\tau_i^{\text{candidate,(e)}}\}_{e=1}^\infty$ is bounded in $[0,1]$ and converges in mean to the oracle threshold τ_i^* as the epoch e increases, i.e.,*

$$\mathbb{E}\Big[\tau_i^{\text{candidate,(e)}}\Big] \xrightarrow{e\to\infty} \tau_i^*. \tag{31}$$

Then the asymmetric EMA threshold $\tau_i^{(e)}$ defined by

$$\tau_i^{(e)} = \begin{cases} \alpha_u\,\tau_i^{\text{candidate,(e)}} + (1-\alpha_u)\tau_i^{(e-1)}, & \tau_i^{\text{candidate,(e)}} > \tau_i^{(e-1)}, \\ \alpha_d\,\tau_i^{\text{candidate,(e)}} + (1-\alpha_d)\tau_i^{(e-1)}, & \tau_i^{\text{candidate,(e)}} \leq \tau_i^{(e-1)}, \end{cases} \tag{32}$$

with $0 < \alpha_d \leq \alpha_u < 1$, satisfies

$$\tau_i^{(e)} \xrightarrow{e\to\infty} \tau_i^* \quad \textit{in mean.} \tag{33}$$

Define the tracking error $e^{(e)} = \tau_i^{(e)} - \tau_i^*$. Let $\alpha^{(e)} \in \{\alpha_u, \alpha_d\}$ denote the update coefficient at epoch e. The EMA recursion gives

$$e^{(e)} = (1-\alpha^{(e)})e^{(e-1)} + \alpha^{(e)}\Big(\tau_i^{\text{candidate,(e)}} - \tau_i^*\Big). \tag{34}$$

Taking expectations and applying the triangle inequality:

$$\Big|\mathbb{E}\Big[e^{(e)}\Big]\Big| \leq (1-\alpha_d)\Big|\mathbb{E}\Big[e^{(e-1)}\Big]\Big| + \alpha_u\Big|\mathbb{E}\Big[\tau_i^{\text{candidate,(e)}} - \tau_i^*\Big]\Big|, \tag{35}$$

where we use $\alpha^{(e)} \geq \alpha_d$ and $\alpha^{(e)} \leq \alpha_u$ to obtain the worst-case bound. Since $(1-\alpha_d) \in (0,1)$, the first term contracts geometrically. The second term vanishes as $e \to \infty$ by assumption. Applying the Gronwall-type inequality recursively yields

$$\left|\mathbb{E}\left[e^{(e)}\right]\right| \leq (1-\alpha_d)^e \left|e^{(0)}\right| + \alpha_u \sum\nolimits_{t=1}^{e} (1-\alpha_d)^{e-t} \varepsilon^{(t)}, \tag{36}$$

where $\varepsilon^{(t)} = |\mathbb{E}[\tau_i^{\text{candidate,(t)}} - \tau_i^*]| \to 0$. Since $(1-\alpha_d) < 1$, both terms converge to zero as $e \to \infty$, establishing $\mathbb{E}[e^{(e)}] \to 0$.

Corollary B.9 (Conservative Noise Suppression) *When $\alpha_d < \alpha_u$, the asymmetric EMA assigns a larger weight to historical information during reliability degradation episodes, i.e., when $\tau_i^{\text{candidate,(e)}} < \tau_i^{(e-1)}$. This asymmetry introduces a conservative bias that slows the propagation of spuriously low reliability signals, thereby reducing the risk of irreversible noisy pseudo-label contamination in CSSL.*

Formally, let $\Delta_u^{(e)} = \tau_i^{\text{candidate,(e)}} - \tau_i^{(e-1)} > 0$ and $\Delta_d^{(e)} = \tau_i^{(e-1)} - \tau_i^{\text{candidate,(e)}} > 0$ denote the upward and downward deviation magnitudes, respectively. Then

$$\tau_i^{(e)} - \tau_i^{(e-1)} = \begin{cases} +\alpha_u \, \Delta_u^{(e)} > +\alpha_d \, \Delta_u^{(e)}, \\ -\alpha_d \, \Delta_d^{(e)} > -\alpha_u \, \Delta_d^{(e)}, \end{cases} \tag{37}$$

showing that threshold decreases are uniformly smaller in magnitude than the symmetric case, while threshold increases are uniformly larger, enabling rapid exploitation of reliability gains without sacrificing noise robustness.

B.4 Reliability-weighted Risk Bound

Finally, we provide a generalization risk bound that quantifies how the proposed reliability calibration reduces the pseudo-label noise level and thereby controls the excess risk of the learned classifier.

Definition B.10 (Pseudo-label Noise Rate) *Let $\mathcal{S}_i$ denote the selected pseudo-label set. Define the pseudo-label noise rate as*

$$\epsilon_i = \frac{1}{|\mathcal{S}_i|} \sum\nolimits_{\mathbf{x}_j \in \mathcal{S}_i} \mathbb{1}[c_j \neq y_j], \tag{38}$$

where c_j and y_j denote the pseudo-label and true label of sample $\mathbf{x}_j$, respectively.

Theorem B.11 (Reliability-Noise Coupling) *Suppose that the GMM posterior satisfies $\gamma_j = P(k_{\text{clean}} \,|\, \widetilde{\mathbf{v}}_j) \geq h_{\min} > 0$ for all $\mathbf{x}_j \in \mathcal{S}_i$. Then the expected pseudo-label noise rate on the selected set satisfies*

$$\mathbb{E}[\epsilon_i] \;\leq\; 1 - \frac{1}{|\mathcal{S}_i|} \sum\nolimits_{\mathbf{x}_j \in \mathcal{S}_i} h_j, \tag{39}$$

where h_j denotes the temporally calibrated posterior reliability of sample $\mathbf{x}_j$.

For any sample $\mathbf{x}_j \in \mathcal{S}_i$, its pseudo-label is correct if and only if the model assigns it to the reliable component. Since $h_j = P(k_{\text{clean}} \,|\, \widetilde{\mathbf{v}}_j)$ is the posterior probability of correctness under the fitted GMM, the expected correctness indicator satisfies

$$\mathbb{E}[\mathbb{1}[c_j = y_j]] \;\geq\; h_j, \tag{40}$$

where the inequality holds because the GMM posterior lower-bounds the true correctness probability under model calibration. Averaging over $\mathcal{S}_i$ and using $\epsilon_i = 1 - |\mathcal{S}_i|^{-1} \sum_j \mathbb{1}[c_j = y_j]$ yields the stated bound.

Corollary B.12 (Threshold-controlled Noise Rate) *Since $\mathcal{S}_i = \{\mathbf{x}_j \,|\, h_j^{(e)} \geq \tau_i^{(e)}\}$, the selection threshold $\tau_i^{(e)}$ directly controls the upper bound on the expected noise rate:*

$$\mathbb{E}[\epsilon_i] \;\leq\; 1 - \tau_i^{(e)}, \tag{41}$$

where the inequality follows from $h_j \geq \tau_i^{(e)}$ for all selected samples. Thus, a higher adaptive threshold $\tau_i^{(e)}$ produces a cleaner pseudo-label set, confirming the noise-suppression mechanism of the proposed adaptive calibration strategy.

Dataset	1%	5%	25%
CIFAR-10	94.97 ± 0.66	95.35 ± 0.57	95.39 ± 0.58
CIFAR-100	76.37 ± 0.47	79.44 ± 0.32	80.36 ± 0.98
ImageNet-R	36.51 ± 2.07	53.96 ± 1.21	57.33 ± 1.37
ImageNet-A	21.32 ± 2.67	27.22 ± 1.92	40.19 ± 2.09

Table 4: Average incremental accuracy (AVG/AIA, %) over three random seeds (1993, 2000, and 2027), reported as mean $\pm$ standard deviation.

Combining Theorems B.8 and B.11, as the epoch index e increases, $\tau_i^{(e)} \to \tau_i^*$ and accordingly $\mathbb{E}[\epsilon_i] \leq 1 - \tau_i^*$. The oracle threshold τ_i^* is determined by the desired reliability level η, providing a direct link between the user-specified quality constraint and the asymptotic noise rate of the selected pseudo-label set.

C Experimental Details

C.1 Datasets and Task Setting.

We evaluate our method on four image-classification benchmarks under a class-incremental CSSL protocol: CIFAR-10, CIFAR-100, ImageNet-R, and ImageNet-A. CIFAR-10 is divided into five tasks of two classes each, CIFAR-100 into ten tasks of ten classes each, and ImageNet-R and ImageNet-A into twenty tasks of ten classes each. The class order is shuffled according to the run seed. We use the standard training and test partitions of the CIFAR datasets. For ImageNet-R, we construct a fixed class-stratified 80/20 training–test split from the original images; for ImageNet-A, which does not provide an official training split, we construct and reuse a fixed class-stratified 80/20 training–test partition using split seed 1993.

For each task, we randomly label 1%, 5%, or 25% of the training images per class, retaining at least one labeled image per class. The remaining training images are treated as unlabeled. After a task is completed, its images are no longer available for training. For direct comparison with the baselines, the main tables and figures report results obtained using seed 1993. We additionally evaluate our complete method using three seeds, i.e., 1993, 2000, and 2027, and report the corresponding mean and standard deviation in Table 4. Unless otherwise specified, baseline and ablation results are obtained from a single run with seed 1993.

C.2 Evaluation Metric.

Following the standard continual-learning protocol Rebuffi et al. (2017), we report Average Incremental Accuracy (AVG). After learning task t, we evaluate the model on the combined test set of all classes seen up to that task. Let A_t denote the resulting Top-1 accuracy. We compute

$$\text{AVG} = \frac{1}{T} \sum_{t=1}^{T} A_t, \tag{42}$$

where T is the total number of tasks. Each A_t is calculated over individual test images, so tasks contribute in proportion to their test-set sizes. Unlike final-stage accuracy, AVG measures performance throughout the entire incremental learning sequence.

C.3 Implementation Details.

All images are resized to 224×224. We initialize a ViT-B/16 backbone with the timm AugReg2 weights pretrained on ImageNet-21K and fine-tuned on ImageNet-1K, and freeze its original parameters. For each task, a separate LoRA expert is inserted into the query and value projections of the last six Transformer blocks, with rank $r = 8$, scaling factor $\alpha = 16$, and dropout rate 0.05. Each task is trained for ten epochs: two labeled-only warm-up epochs followed by eight semi-supervised epochs. We use AdamW with a learning rate of 5×10^{-4} and weight decay of 10^{-4}. Each epoch contains 150 iterations, with 16 labeled and 48 unlabeled images per iteration. Strong augmentation uses RandAugment with two operations and magnitude nine.

C.4 Pseudo-label Selection and Optimization.

At each semi-supervised epoch, the GMM is calibrated using every unique unlabeled image of the current task once, irrespective of repeated mini-batch sampling during training. The selection temperature is 2, whereas hard pseudo-label cross-entropy uses a student temperature of 1. The target expected posterior reliability η is set to 0.95 for the 1% labeling ratio and 0.90 for the 5% and 25% ratios. The two-component GMM uses full covariance matrices, at most 25 EM iterations, a convergence tolerance of 10^{-4}, and covariance shrinkage of 0.2. Its three initialization quantiles are $\{0.10, 0.25, 0.35\}$.

The sample-wise posterior EMA uses momentum $m_h = 0.7$. A candidate selection threshold is adopted immediately when it increases; when it decreases, the updated threshold is 0.7 times the previous threshold plus 0.3 times the candidate. No additional hard selection gates are applied. Selected images provide hard pseudo-label cross-entropy on their strongly augmented views, and the unlabeled loss is normalized by the number of selected images rather than the full unlabeled batch size. If no valid GMM is available, the previous valid state is reused when possible; otherwise, that epoch uses labeled supervision only.

C.5 Baseline and State-of-the-Art Configurations

All comparison methods follow the common protocol described in Secs. C.1–C.4. We preserve their original continual-learning objectives and modify only the data interface required by the shared CSSL setting. Method-specific configurations are summarized in Table 5.

The original optimization settings and loss coefficients are retained for each method. Ground-truth labels of unlabeled samples are used only for offline diagnostic statistics and never affect training or sample selection.

Method	Pseudo-label configuration	Retained mechanism
Conventional continual-learning baselines		
iCaRL	Confidence ($\tau = 0.95$)	Classification and distillation
FOSTER	Confidence ($\tau = 0.95$)	Feature boosting and compression
DER	Confidence ($\tau = 0.95$)	Expandable representation
DER++	Confidence ($\tau = 0.50$)	Label and logit replay
Recent CSSL and parameter-efficient approaches		
DSGD	Progressive pseudo-labeling	Dynamic sub-graph distillation
Keep-LoRA	FixMatch ($\tau = 0.95$)	16-dim. LoRA with subspace retention
USP	Native selector ($\tau = 0.95$)	Divide-and-conquer learning
TACLE	Task-wise FixMatch ($\tau_0 = 0.90$)	Class-aware calibration
CoVar	Confidence and covariance weighting	Covariance-aware consistency

Table 5: Method-specific configurations of the comparison approaches. Other settings follow the common protocol in Secs. C.1–C.4.

C.6 GMM Validity Check

A newly estimated GMM is accepted only when its partition satisfies several validity conditions. First, at least 64 unique unlabeled samples with finite predictive-semantic coordinates must be available. Following the component identification rule introduced in the main text, let $(\pi_{\mathrm{r}}, \boldsymbol{\mu}_{\mathrm{r}}, \boldsymbol{\Sigma}_{\mathrm{r}})$ and $(\pi_{\mathrm{n}}, \boldsymbol{\mu}_{\mathrm{n}}, \boldsymbol{\Sigma}_{\mathrm{n}})$ denote the mixture weight, mean vector, and covariance matrix of the relatively reliable and unreliable components, respectively. Their centers in the standardized predictive-semantic reliability space are represented as $\boldsymbol{\mu}_{\mathrm{r}} = (\mu_{\mathrm{r}}^{p}, \mu_{\mathrm{r}}^{q})$ and $\boldsymbol{\mu}_{\mathrm{n}} = (\mu_{\mathrm{n}}^{p}, \mu_{\mathrm{n}}^{q})$. The subscripts r and n describe relative reliability modes rather than ground-truth correct and incorrect pseudo-labels.

Each component must account for at least 5% of the total mixture mass:

$$\min(\pi_{\mathrm{r}}, \pi_{\mathrm{n}}) \geq 0.05. \tag{43}$$

The center of the relatively reliable component must also be no lower than that of the other component along either reliability coordinate:

$$\mu_{\mathrm{r}}^{p} \geq \mu_{\mathrm{n}}^{p}, \qquad \mu_{\mathrm{r}}^{q} \geq \mu_{\mathrm{n}}^{q}. \tag{44}$$

We further evaluate component separability using the pooled-covariance Mahalanobis distance:

$$D_{\mathrm{sep}} = \sqrt{(\boldsymbol{\mu}_{\mathrm{r}} - \boldsymbol{\mu}_{\mathrm{n}})^{\top} \left[\frac{\boldsymbol{\Sigma}_{\mathrm{r}} + \boldsymbol{\Sigma}_{\mathrm{n}}}{2}\right]^{-1} (\boldsymbol{\mu}_{\mathrm{r}} - \boldsymbol{\mu}_{\mathrm{n}})}. \tag{45}$$

The fitted mixture is considered sufficiently separated only when $D_{\mathrm{sep}} \geq 0.5$. In addition, posterior-based threshold calibration must produce at least eight selected samples while satisfying the target reliability η defined in Sec. C.4. Among all valid multi-start solutions, we retain the one with the largest average log-likelihood. If none of the candidate solutions satisfies these conditions, the fallback strategy described in Sec. C.4 is applied.

C.7 Experimental Setup for Main-Text Visualizations

Common Protocol. Unless otherwise specified, all visualizations are produced on ImageNet-R using seed 1993 and the same data split, class order, and optimization settings described in Sections C.1–C.4. Pseudo-label precision and selection ratios are computed over unique unlabeled images, such that repeated minibatch sampling does not duplicate a sample in the statistics. Ground-truth labels are used only for post-hoc evaluation and visualization, and are never used for pseudo-label selection or model optimization.

Introduction Visualizations. For Fig. 1(a), we use the unique unlabeled samples recorded at the final semi-supervised epoch of every task. At each selection ratio $r \in \{20\%, 40\%, 60\%, 80\%, 90\%\}$, every method retains the same number $K_r = \max(1, \lfloor rN_u \rfloor)$ of samples within each task according to its native selection score. RCF ranks samples using the temporally calibrated posterior, whereas CL-LoRA and USP use their original confidence-based selection scores. Pseudo-label precision is micro-averaged over all selected samples from the 20-task stream. For Fig. 1(b), we replace the original confidence selector of CL-LoRA with our two-dimensional GMM selector while keeping its architecture, optimization procedure, and task-ID inference protocol unchanged. Figure 1(c) pools the unique unlabeled samples recorded at the final semi-supervised epoch of all 20 tasks. The predictive and semantic coordinates are independently divided into five equal-frequency groups over the pooled samples, and each cell reports the pseudo-label precision in the corresponding two-dimensional bin. Figure 1(d) visualizes the unique unlabeled samples from the final semi-supervised epoch of Task 4. Point colors are determined only by post-hoc pseudo-label correctness. The prediction-only boundary is chosen to retain the same number of samples as the two-dimensional GMM, resulting in a matched-K comparison.

Ablation and Analytical Visualizations. Figures 3(a) and 3(b) report AIA over the complete 20-task sequence under the 1%, 5%, and 25% labeling ratios using seed 1993. The coordinate ablation retains predictive reliability, semantic reliability, or their joint representation, whereas the standardization ablation changes only whether robust coordinate standardization is applied before GMM fitting. All non-target components are kept unchanged. For Fig. 3(c), we pool the unique unlabeled samples exported at the final semi-supervised epoch of all 20 tasks and rank them globally using raw confidence S, predictive reliability p, the current-epoch GMM posterior γ, and the temporally calibrated posterior h. The panel reports pseudo-label precision at matched coverage levels of 20%, 40%, 60%, 80%, and 100%. Figure 3(d) visualizes the reliability space at the final semi-supervised epoch of Task 1 under the 1% labeling ratio. The one-dimensional, shared-covariance linear GMM, and full-covariance curved GMM boundaries are displayed under the same snapshot-level selection budget. The annotated AIA and weighted pseudo-label precision are obtained from their respective complete 20-task runs rather than from this single snapshot.

C.8 Routing Implementation

Ground-truth task identities are unavailable during inference. After learning each task, we construct a label-free router from the frozen ViT features of all unique training images in that task. Cosine

Selection Boundary	1% Labels			5% Labels			25% Labels		
	AIA	PL Acc.	Sel.	AIA	PL Acc.	Sel.	AIA	PL Acc.	Sel.
1D predictive	29.77	64.0	26.9	52.94	93.7	56.5	**57.80**	98.5	71.1
2D linear (shared cov.)	**34.85**	**76.1**	24.7	53.13	**98.0**	40.7	57.62	**99.5**	56.6
2D curved (full cov.)	34.30	71.9	33.4	**53.48**	94.8	62.0	57.71	98.6	72.6

Table 6: Detailed reliability-boundary comparison on ImageNet-R. PL Acc. and Sel. denote weighted pseudo-label accuracy and the selected pseudo-label ratio, respectively. All values are reported in %.

Scale Floor β	1% Labels	5% Labels	25% Labels
0.01	34.12	53.59	57.73
0.05	**34.30**	53.48	57.71
0.10	33.84	**53.65**	**57.81**

Table 7: Sensitivity to the robust-scale floor β on ImageNet-R. AIA (%) is reported using seed 1993.

K-means first produces ten core prototypes. Within each core cluster, samples whose distances to the cluster center lie between the 0.70 and 0.95 quantiles are further represented by up to three boundary prototypes. Neither class labels nor pseudo-labels are used during this process. Cosine K-means runs for at most 25 iterations with a convergence tolerance of 10^{-4}.

Let $\mathbf{h}$ denote the normalized frozen-ViT feature of a test image and let $\mathbf{P}_i$ contain the normalized core and boundary prototypes of task $\mathcal{T}_i$. We compute the routing decision as

$$\mathbf{s}_i = \mathbf{P}_i\mathbf{h}, \qquad r_i = \mathrm{MeanTopK}(\mathbf{s}_i, 3), \qquad \widehat{i} = \arg\max_i r_i, \tag{46}$$

where $\mathrm{MeanTopK}(\mathbf{s}_i, 3)$ returns the mean of the three largest prototype similarities. Only the expert associated with the selected task $\widehat{i}$ is executed for classification. Therefore, routing errors are directly reflected in the reported classification accuracy.

D Additional Experiment Results

D.1 Detailed Results of Different Reliability Partitioning Boundaries

Figure 3(d) visualizes the boundary comparison under the most challenging 1% labeling ratio. Here, we report the complete results under all three labeling ratios. The one-dimensional variant uses predictive reliability alone, while both two-dimensional variants jointly model predictive and semantic evidence. The linear GMM shares one covariance matrix between its two components, whereas the curved GMM uses component-specific full covariance matrices.

Beyond the 1% setting discussed in the main paper, Table 6 reveals how the three boundaries behave as labeled supervision increases. At 5% labels, the linear and curved GMMs raise pseudo-label accuracy from 93.7% to 98.0% and 94.8%, respectively, while selecting fewer samples than the one-dimensional variant. Their corresponding AIA values remain higher at 53.13% and 53.48%, compared with 52.94%. At 25% labels, all variants exceed 98% pseudo-label accuracy and their AIA differences remain within 0.18 points, although the two-dimensional variants continue to retain fewer but more accurate pseudo-labels. These results show that semantic evidence mainly improves the reliability and efficiency of selection, whereas its downstream advantage naturally decreases when stronger labeled supervision is available.

D.2 Hyperparameter Sensitivity

Robust-Scale Floor for GMM Fitting. During robust coordinate standardization, the scale floor β prevents an unusually small median absolute deviation from excessively amplifying minor coordinate variations. We evaluate $\beta \in \{0.01, 0.05, 0.10\}$ while keeping all other settings fixed.

As shown in Table 7, performance remains stable across the three tested values. Moreover, neither a smaller nor a larger floor consistently performs best across labeling ratios. We therefore use the

η	Labeling ratio		
	1%	5%	25%
0.85	33.65	**53.64**	**57.72**
0.90	33.60	<u>53.48</u>	<u>57.71</u>
0.95	**<u>34.30</u>**	53.54	57.63

Table 8: Sensitivity to the desired reliability level η on ImageNet-R. Average incremental accuracy (AVG, %) is reported. Bold values denote the best result under each labeling ratio, and underlined values indicate the settings used in the main experiments.

Component	**Coefficient**	**1%**	**5%**	**25%**
Posterior EMA	$m_h = 0.0$	33.21	**53.60**	57.67
	$m_h = 0.7$	**<u>34.30</u>**	<u>53.48</u>	**<u>57.71</u>**
	$m_h = 0.9$	33.38	52.55	57.63
Threshold update	$\alpha_d = 1.0$	33.64	53.36	57.42
	$\alpha_d = 0.3$	**<u>34.30</u>**	**<u>53.48</u>**	**<u>57.71</u>**
	$\alpha_d = 0.1$	33.30	53.21	57.52

Table 9: Sensitivity to the temporal calibration parameters on ImageNet-R. Each parameter is varied independently, and AVG (%) is reported using seed 1993. Bold values denote the best result within each parameter group, while underlined values indicate the settings used in the main experiments.

intermediate value $\beta = 0.05$, which prevents near-zero robust scales without imposing an unnecessarily large lower bound.

Desired Reliability Level η. The desired reliability level η in Eq. (19) determines the largest posterior-ranked subset whose mean temporally calibrated reliability remains no lower than η. Increasing η therefore imposes a stricter reliability requirement, whereas decreasing it allows more pseudo-labels to participate in training. We evaluate $\eta \in \{0.85, 0.90, 0.95\}$ on ImageNet-R.

As shown in Table 8, the 1% setting benefits from the stricter reliability requirement $\eta = 0.95$, confirming that conservative pseudo-label selection is particularly important under extreme label scarcity. With 5% and 25% labels, the performance ranges are only 0.16 and 0.09 percentage points, respectively, indicating limited sensitivity to η when stronger labeled supervision is available. We therefore use $\eta = 0.95$ for 1% labels and $\eta = 0.90$ for 5% and 25% labels, retaining a moderately conservative reliability constraint without excessively restricting pseudo-label coverage.

Posterior and Threshold Calibration Parameters. The temporal calibration mechanism contains two complementary components. The posterior momentum m_h in Eq. (17) controls the temporal aggregation of sample-level reliability, while α_u and α_d in Eq. (21) control the asymmetric update of the selection threshold. We independently evaluate m_h and α_d, with $\alpha_u = 1$ fixed so that threshold increases are adopted immediately, whereas threshold decreases are smoothed.

As shown in Table 9, $m_h = 0.7$ achieves the best performance under the 1% and 25% labeling ratios and remains close to the best result at 5%. A larger value of $m_h = 0.9$ retains too much historical information and responds more slowly to the evolving reliability distribution. For threshold calibration, $\alpha_d = 0.3$ consistently performs best across all three labeling ratios, providing a balance between temporal stability and adaptation. We therefore use $m_h = 0.7$, $\alpha_u = 1$, and $\alpha_d = 0.3$ in the main experiments.

Selection and Hard-CE Temperatures. The selection temperature controls the weak-view probability distributions used to construct the predictive reliability coordinate, whereas the hard-CE student temperature scales the strong-view logits in pseudo-label optimization. We analyze the two temperatures separately: the hard-CE temperature is fixed to 1 when varying the selection temperature, and the selection temperature is fixed to 2 when varying the hard-CE temperature.

As shown in Table 10, the selection temperature has a relatively moderate influence on performance. A value of 2 achieves the highest AIA under the more challenging 1% labeling ratio and remains

Parameter	Value	1%	5%
Selection temperature	1.0	33.31	**54.01**
	1.5	34.01	53.65
	2.0	**34.30**	53.48
Hard-CE student temperature	1.0	**34.30**	**53.48**
	1.5	32.18	52.60

Table 10: Sensitivity to the selection and hard-CE student temperatures on ImageNet-R. Each temperature is varied independently, and AIA (%) is reported using seed 1993. Bold values denote the best result within each parameter group, while underlined values indicate the settings used in the main experiments.

Label ratio	$r_{\mathrm{LoRA}} = 2$	$r_{\mathrm{LoRA}} = 4$	$r_{\mathrm{LoRA}} = 8$	$r_{\mathrm{LoRA}} = 16$
1%	**35.56**	34.90	33.99	33.67
5%	53.38	53.47	**53.48**	52.94

Table 11: Sensitivity to the LoRA rank on ImageNet-R. Results report AIA (%).

within 0.53 points of the best result at 5%. In contrast, increasing the hard-CE student temperature from 1 to 1.5 consistently degrades performance, indicating that softening the strong-view logits weakens hard pseudo-label supervision. We therefore use a selection temperature of 2 and an unscaled hard-CE student temperature of 1 in the main experiments.

LoRA Rank. We evaluate the sensitivity of RCF to the LoRA rank $r_{\mathrm{LoRA}} \in \{2, 4, 8, 16\}$ on ImageNet-R with 1% and 5% labels, while keeping all other settings fixed. As shown in Table 11, the AIA variation is limited to 1.89 percentage points with 1% labels and 0.54 points with 5% labels. Moreover, increasing the rank does not consistently improve performance, indicating that the effectiveness of RCF does not depend on a large adapter capacity. We retain $r_{\mathrm{LoRA}} = 8$ as the moderate, pre-specified configuration used in all main experiments, rather than selecting a different rank for each labeling ratio or dataset.

D.3 Generality of RCF as a Plug-in Pseudo-Label Selector

To examine whether the proposed reliability estimator is tied to the RCF architecture, we replace the original confidence-based selector in CL-LoRA with our predictive–semantic GMM selector while keeping its backbone, optimization procedure, and task-ID inference protocol unchanged.

As shown in Table 12, replacing confidence-based selection with our selector substantially improves pseudo-label precision under both labeling ratios. With 1% labels, pseudo-label precision increases from 47.20% to 83.87%, accompanied by improvements in final-stage average accuracy from 47.20% to 59.93% and in AIA from 48.10% to 61.44%. With 5% labels, our selector increases pseudo-label precision from 88.26% to 97.68% while reducing the selection ratio from 96.47% to 52.27%. Meanwhile, it maintains comparable final-stage average accuracy (88.68% vs. 88.72%) and AIA (89.40% vs. 89.67%). These results demonstrate that prioritizing pseudo-label reliability can provide stronger supervision under scarce labels and preserve competitive performance when labeled supervision is more sufficient. They also demonstrate that the proposed selector can be transferred to another continual-learning architecture.

D.4 Further Calibration of Prototype-Based Semantic Evidence

The preceding ablations show that prototype-based semantic evidence provides useful information complementary to predictive reliability. We further investigate whether its contribution can be strengthened through class-adaptive calibration. Since prototype quality may vary across categories, particularly under limited supervision, we introduce a category-level reliability factor r_c to control the contribution of semantic evidence.

Labels	Selector	PL Prec.	Selected	Final-stage Acc.	AIA
1%	Confidence	47.20	96.41	47.20	48.10
	Ours	**83.87**	34.29	**59.93**	**61.44**
5%	Confidence	88.26	96.47	**88.72**	**89.67**
	Ours	**97.68**	52.27	88.68	89.40

Table 12: Plug-in evaluation on CL-LoRA using ImageNet-R. PL precision and selection ratio are computed over unique unlabeled samples. All model results use task-ID inference.

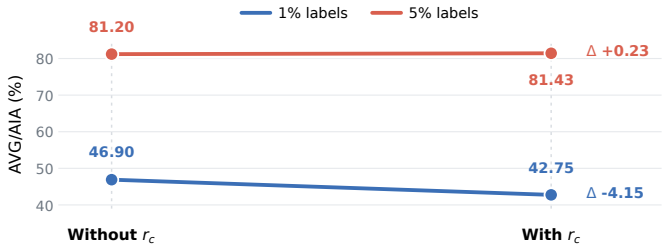


Figure 4: Effect of category-level reliability calibration r_c on ImageNet-R under task-ID-based inference. Each line connects paired runs with otherwise identical configurations. Values report AVG/AIA (%).

For a sample with pseudo-label $\widehat{y}_j$, the calibrated posterior is obtained by interpolating the predictive–semantic posterior γ_j^{xy} and the prediction-only posterior γ_j^{x}:

$$\bar{\gamma}_j = r_{\widehat{y}_j}^{\alpha_r} \gamma_j^{xy} + \left(1 - r_{\widehat{y}_j}^{\alpha_r}\right) \gamma_j^{x}, \tag{47}$$

where α_r controls the calibration strength. This design is intended to preserve reliable semantic evidence while reducing the influence of potentially inaccurate class prototypes.

To evaluate this hypothesis, we compare the calibrated posterior with the original two-dimensional posterior under otherwise identical settings.

Contrary to our expectation, Figure 4 shows that r_c does not provide a consistent improvement. With 1% labels, enabling r_c decreases AVG/AIA from 46.90% to 42.75%, while with 5% labels it only changes AVG/AIA from 81.20% to 81.43%. This suggests that the class-level calibration may introduce additional uncertainty under limited supervision and unnecessarily suppress useful semantic evidence.

We therefore exclude r_c from the final framework and directly use the two-dimensional posterior γ_j^{xy} for subsequent calibration and pseudo-label selection.

D.5 From Pseudo-Label Reliability to Training Utility

The main paper establishes a reliability-first principle: unreliable pseudo-labels introduce misleading supervision, whereas rejecting an unlabeled sample only removes a potentially useful training signal. However, pseudo-label reliability is a prerequisite rather than the sole determinant of effective semi-supervised learning. After sufficiently reliable pseudo-labels have been identified, their contribution also depends on whether the selected samples provide informative and complementary supervision. We therefore conduct a controlled analysis to examine how pseudo-label reliability is translated into downstream model performance.

Specifically, we construct four settings on ImageNet-R with 5% labels under task-ID-based inference. Setting A uses the pseudo-labels selected by our complete method. Setting B keeps exactly the same selected samples but replaces their pseudo-labels with oracle labels, isolating the influence of residual pseudo-label errors. Setting C uses a count- and class-matched random subset with oracle labels, allowing us to examine the training utility of the selected samples under identical supervision quantity and label correctness. Setting D further uses all unlabeled samples with oracle labels to estimate the additional benefit brought by supervision quantity.

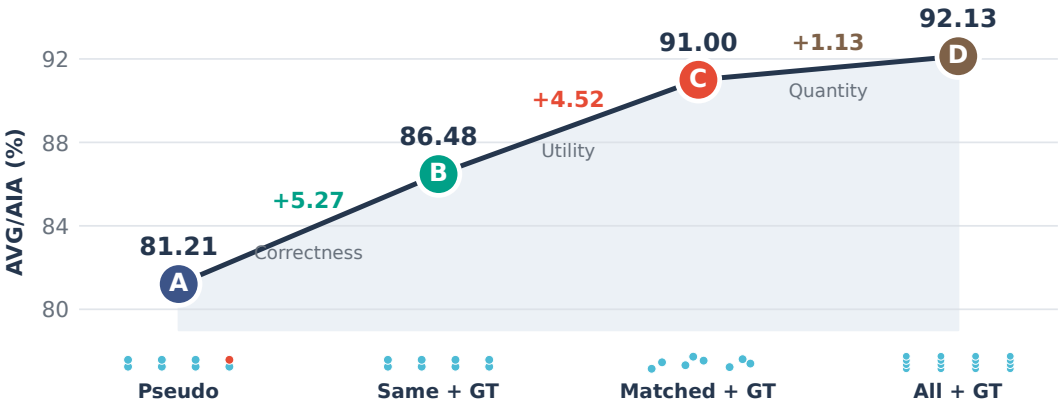


Figure 5: From pseudo-label reliability to training utility on ImageNet-R with 5% labels. The four controlled settings progressively isolate the effects of label correctness, sample utility, and supervision quantity.

Method	Trainable/task (M)	Params (M)	FLOPs (G)	Inference form
TACLE	0.008	85.952	35.14	Single pass
USP	0.314	86.113	33.70	Single pass
Keep-LoRA	0.692	86.884	35.61	Single pass
CL-LoRA	0.438	96.551	≈700.0†	All experts
CoVar	0.155	88.902	50.50	Route + one expert
RCF (Ours)	**0.155**	**88.902**	**50.50**	Route + one expert

†CL-LoRA evaluates all stored adapter branches during task-free inference. Its FLOPs are estimated at the final 20-task stage and therefore increase with the number of learned tasks. CoVar uses the same expert and routing architecture as RCF under our unified comparison protocol.

Table 13: Architecture-level computational cost at the end of the 20-task ImageNet-R sequence. FLOPs are computed for one test image using a common operator-profiler. The CL-LoRA FLOPs are analytically estimated from its native all-expert inference process.

As shown in Fig. 5, replacing the selected pseudo-labels with oracle labels improves AVG/AIA from 81.21% to 86.48%. This 5.27-point gain confirms the main-paper observation that pseudo-label reliability must be secured before increasing the amount of unlabeled supervision. Under the same sample count, class distribution, and oracle supervision, the matched random subset further improves AVG/AIA to 91.00%. This result indicates that the reliable samples selected by the current estimator are not always the most informative for model optimization and may contain easy or redundant instances. Finally, using all oracle-labeled unlabeled samples yields a comparatively smaller improvement of 1.13 points.

These results complement, rather than contradict, the reliability-first principle. Reliable pseudo-labels provide the necessary foundation for safe continual optimization, while sample diversity and training utility determine how effectively this reliable supervision is converted into model accuracy. Accordingly, a promising extension is to preserve the reliability guarantee of RCF while favoring more informative samples within the reliable candidate set.

D.6 Computational Cost and Efficiency

We compare the computational cost of RCF with representative CSSL and parameter-efficient continual-learning methods under the same ViT-B/16 backbone. The comparison is conducted at the end of the 20-task ImageNet-R sequence with an input resolution of 224×224. Model parameters exclude stored training samples, while the reported trainable parameters correspond to the parameters optimized when learning the current task.

As shown in Table 13, RCF optimizes only 0.155M parameters for the current task, corresponding to 0.181% of the frozen ViT backbone. This is approximately half the trainable parameter count of

USP, 35% of that of CL-LoRA, and 22% of that of Keep-LoRA. After all 20 tasks, the task-specific experts introduce 3.103M parameters, or 3.617% of the backbone, resulting in a total model size of 88.902M parameters. The prototype router additionally requires only 2.344 MiB of storage.

RCF incurs higher inference cost than single-pass methods because it first extracts frozen semantic features for routing and then executes the selected expert. Its 50.50 GFLOPs correspond to approximately 1.4–$1.5\times$ the cost of TACLE, USP, and Keep-LoRA. However, RCF executes only one selected expert, and its dominant neural-network computation remains independent of the number of learned tasks. Although prototype comparison grows with the number of tasks, it contributes only about 0.0012 GFLOPs at the final 20-task stage. In contrast, CL-LoRA evaluates every stored expert and requires approximately 700 GFLOPs at the final stage, about $13.9\times$ the inference cost of RCF.

We further measure the absolute runtime of RCF on an NVIDIA RTX 3090. Routed inference requires 192.99 ms for a single image and achieves 65.01 images per second with a batch size of 64. A complete semi-supervised optimization step processes 38.21 images per second and uses 4.07 GiB of peak allocated GPU memory. The complete 20-task ImageNet-R experiment with 5% labels requires approximately 5.94 hours.

The additional probabilistic calibration introduces limited overhead. For 1,200 unique unlabeled samples, GMM fitting, posterior calibration, and threshold estimation require only 0.102 seconds, whereas the complete calibration process requires 26.75 seconds. Therefore, GMM estimation accounts for approximately 0.38% of the calibration time; most of the additional cost arises from extracting predictive and frozen semantic features rather than from the proposed probabilistic modeling itself.