Mitigating New-Tail Collapse in Long-Tailed Class-Incremental Learning with CLIP

Hyeokjun Lee[1], **Sungmin Cha**[2*], **YoungJoon Yoo**[1†]
[1]Department of Artificial Intelligence, Chung-Ang University
[2]Meta Superintelligence Labs
{gurwns8926, yjyoo3312}@cau.ac.kr
sungmincha@meta.com

Abstract

Pre-trained vision–language models such as CLIP provide strong priors for class-incremental learning (CIL), yet they degrade sharply under long-tailed class streams. We show that this degradation is not merely a byproduct of forgetting: in frozen-CLIP long-tailed CIL, newly introduced tail classes are disproportionately absorbed into previously learned old classes, a boundary-level failure we call *new-tail collapse*. We diagnose this failure with an old-new margin and analyze it through a boundary-risk view, where CLIP semantic overlap, sparse tail supervision, and old-class expansion jointly increase old-dominant violations. Motivated by this view, we propose Smooth Boundary-Risk Loss (SBRL), which estimates effective old rivals from CLIP's vision-language geometry and suppresses their smooth old-side pressure against new-tail samples. MC-SMERG further consolidates task-wise adapters into a single efficient model. Across ImageNetSubset-LT, CIFAR100-LT, and Food101-LT, our method consistently outperforms strong CLIP-based CIL baselines and substantially reduces boundary violations in high-similarity, low-shot regimes.

1 Introduction

Pre-trained vision–language models (VLMs) such as CLIP (Radford et al., 2021) have become strong backbones for class-incremental learning (CIL), enabling continual recognition with minimal task-specific supervision. However, real-world visual streams are often long-tailed, where newly arriving classes contain only a few examples (Liu et al., 2019; Kang et al., 2020). Under this regime, CLIP-based CIL degrades sharply, and the usual forgetting-centric explanation is incomplete. Existing methods largely focus on preserving old knowledge through replay (Rebuffi et al., 2017), distillation (Li and Hoiem, 2016; Kang et al., 2022), or reweighting (He, 2024; Cui et al., 2019), but they do not directly address how newly introduced tail classes fail to establish their own decision regions—the failure mode we identify next.

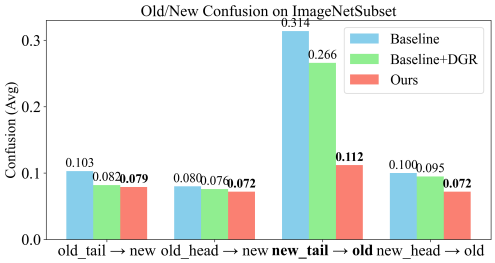


Figure 1: Old/New confusion on ImageNetSubset. We report Avg/Last over Tasks 1–4. DGR denotes Gradient Reweighting (He, 2024), an existing LT-CIL method. Our method substantially reduces the new-tail→old collapse while also improving old→new confusions.

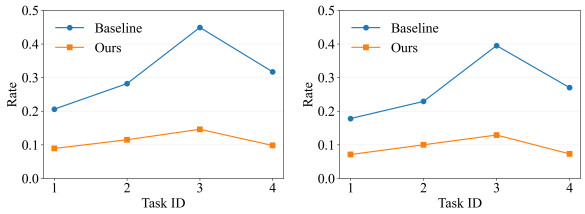
We identify a highly asymmetric failure mode in frozen-CLIP long-tailed CIL: newly introduced tail classes are disproportionately absorbed into previously learned old classes. We call this phenomenon **new-tail collapse**. As shown in Figure 1, on ImageNetSubset-200 constructed from ImageNet (Russakovsky et al., 2015) under a 5-task class-incremental protocol, new-tail→old confusion is substantially larger than new-head→old confusion. This asymmetric absorption suggests that the degradation cannot be fully explained by old-class forgetting alone; instead, the newly introduced tail classes appear to be selectively overruled by previously learned classes.

To test whether this confusion reflects a boundary-level failure, we introduce an old-new boundary diagnostic based on the gap between the largest old-class logit and the largest current-task new-class logit. A positive margin means that some old class outranks all current new classes, so the

*This work, including all code, was conducted by the authors in their personal capacity (co-authors at Chung-Ang University) and does not represent the views of Meta.

†Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 41837–41856
October 24-29, 2026 ©2026 Association for Computational Linguistics



(a) frac_pos($\Delta > 0$) (b) frac_big ($\Delta > 0.5$)

Figure 2: Quantifying decision-boundary collapse for new-tail classes. We track the margin $\Delta = \max(\text{old logit}) - \max(\text{new logit})$ on *new-tail* samples across incremental tasks, and report the fractions with *frac_pos* and *frac_big*.

sample lies on the old side of the old-vs-new decision boundary. As shown in Figure 2 under the same protocol, the violation rates frac_pos and frac_big reveal that baseline models increasingly absorb new-tail samples into existing old-class regions as tasks progress, whereas our method substantially reduces these violations.

We analyze this behavior through a boundary-risk view: CLIP semantic overlap, scarce tail supervision, and the growing old-class pool jointly increase the chance that an old class dominates a new-tail sample. We formalize this mechanism in Section 3.2.

Motivated by this view, we propose a CLIP-specific boundary-risk minimization framework. Rather than treating boundary collapse only as a diagnostic, we turn it into a training signal. We first estimate effective old rivals using CLIP's shared vision-language space, where text prototypes provide stable semantic references and adapter-aware image prototypes provide distribution-sensitive evidence. We then introduce **Smooth Boundary-Risk Loss** (SBRL), which suppresses the smooth old-side pressure from multiple effective old rivals and enforces the true new class to exceed it. Finally, MC-SMERG consolidates task-wise adapters into a single efficient model through lightweight mode-connectivity-based merging.

Experiments on ImageNetSubset-LT, CIFAR100-LT, and Food101-LT show consistent gains over strong CLIP-based CIL baselines while remaining exemplar-free. Beyond accuracy, our boundary analysis shows that the proposed method substantially reduces old-dominant violations, especially in the high-similarity and low-shot regimes predicted by our boundary-risk view.

Our contributions are:

- We identify **new-tail collapse** as a boundary-level failure mode in frozen-CLIP LTCIL and analyze its drivers through a boundary-risk view.
- We propose **Smooth Boundary-Risk Loss** (SBRL), which directly converts the diagnosed old-over-new boundary event into a CLIP-guided, multi-rival, pairwise risk-adaptive objective.
- We demonstrate consistent gains across three long-tailed benchmarks and show that violation reduction concentrates in the predicted high-similarity, low-shot regime.

2 Related Work

2.1 Class-Incremental Learning

Class-incremental learning (CIL) aims to learn novel classes sequentially while maintaining performance on previously learned ones. A central challenge is catastrophic forgetting, where learning new data induces abrupt degradation on old classes (McCloskey and Cohen, 1989). Prior work broadly mitigates forgetting via regularization and distillation, which constrain updates to preserve previous knowledge (Kirkpatrick et al., 2017; Li and Hoiem, 2016; Kang et al., 2022), and replay, which revisits past information through stored exemplars or generated samples during training (Rebuffi et al., 2017; Castro et al., 2018; Wu et al., 2019; Shin et al., 2017). While exemplar replay can be effective, it introduces memory, compute, and privacy constraints, motivating lightweight or exemplar-free alternatives. Recent PTM-based CIL methods also explore shared adapter tuning and representation alignment; RSIAT (Zhao et al., 2026), for example, maintains a single adapter in a visual-only PTM-CIL setting with a learned cosine classifier.

2.2 CIL with CLIP

Pre-trained VLMs such as CLIP (Radford et al., 2021) have become strong backbones for CIL, where a frozen encoder already offers non-trivial robustness to forgetting (Thengane et al., 2022). Most CLIP-CIL methods therefore pursue parameter-efficient adaptation, either on the classifier/prompt side via text-image semantics and calibration, or on the visual side via lightweight modules such as adapters, projections, or probabilistic fine-tuning (Huang et al., 2024; Zhou et al., 2025b; Yu

et al., 2024; Jha et al., 2024). Several pipelines stabilize old knowledge via feature-space replay using class-wise statistics rather than raw exemplars (Huang et al., 2024; Hayes and Kanan, 2020; Tang et al., 2023; Zhang et al., 2023). Recent CLIP-CIL works also explore external knowledge injection or modality-gap preservation (Zhou et al., 2025a; Huang et al., 2025). These directions are complementary to ours, which focuses on local old-new boundary risk under long-tailed supervision.

2.3 Long-Tailed CIL and Boundary Diagnostics

Real-world continual streams are often long-tailed, where head classes dominate while tail classes have scarce samples (Liu et al., 2019). Beyond static long-tailed recognition remedies (Kang et al., 2020; Cui et al., 2019), recent long-tailed CIL (LTCIL) work studies imbalance-aware training to better preserve tail performance under sequential updates, including gradient/loss reweighting and geometry/statistics-based stabilization (He, 2024; Lai et al., 2025; Hu et al., 2025). Dynamic adaptation has also been explored via task-imbalanced prompting and dynamically tuned adapters to modulate capacity according to imbalance (Hong et al., 2024; Gu et al., 2025). More recently, AdaPrior (Bhat et al., 2026) formulates LTCIL as dynamic model-prior misalignment and performs online prior estimation for training-time and post-hoc logit correction. While these approaches address global imbalance, representation drift, or prior bias, less attention has been given to the sample-level mechanism by which specific old classes absorb new-tail samples at the decision boundary. Our work fills this gap by diagnosing new-tail collapse at the logit-boundary level and turning the corresponding old-over-new event into a trainable objective with SBRL.

3 Method

Figure 3 gives an overview of our framework. At each task, we estimate effective old rivals using CLIP's vision-language geometry, suppress their old-side pressure with SBRL, and consolidate the resulting adapter with MC-SMERG. We next formalize the boundary-risk view and describe each component.

3.1 Preliminaries: Frozen-CLIP Long-Tailed CIL

We consider class-incremental learning under a long-tailed class distribution. A learner receives a sequence of tasks $\{\mathcal{D}_t\}_{t=0}^{T-1}$, where each task $\mathcal{D}_t = \{(x_i, y_i)\}$ contains samples from a disjoint class set $\mathcal{C}_t$. After training on task t, the model is evaluated over the cumulative label space $\mathcal{C}_{\le t} = \bigcup_{k=0}^{t} \mathcal{C}_k$ without access to task identity. For $t > 0$, we denote previously learned classes by $\mathcal{C}_{<t} = \bigcup_{k=0}^{t-1} \mathcal{C}_k$ and newly introduced classes by $\mathcal{C}_t$.

We adopt a frozen-CLIP adaptation setting. Let $\phi(\cdot)$ be the frozen CLIP image encoder and $\psi(\cdot)$ be the frozen CLIP text encoder. Only a lightweight adapter $A(\cdot)$ is trained on top of the frozen image representation. Given an image x, we compute the normalized adapted image feature

$$z(x) = \mathrm{norm}\left(A(\phi(x))\right). \tag{1}$$

For each class $c \in \mathcal{C}_{\le t}$, we obtain a text prototype

$$w_c = \mathrm{norm}\left(\psi(p(c))\right), \tag{2}$$

where $p(c)$ is a prompt template containing the class name. Prediction is performed using cosine-similarity logits:

$$s_c(x) = \tau z(x)^\top w_c, \quad \hat{y}(x) = \arg\max_{c \in \mathcal{C}_{\le t}} s_c(x), \tag{3}$$

where $\tau > 0$ is the CLIP logit scale.

This setting preserves CLIP's pretrained vision-language geometry while enabling parameter-efficient adaptation. However, under long-tailed task streams, newly introduced tail classes receive sparse supervision while the old-class pool grows over time. We next show how this induces a boundary-level failure beyond conventional forgetting.

3.2 Boundary-Risk View of New-Tail Collapse

We formalize new-tail collapse as an old-dominant boundary failure. At task $t > 0$, let $\mathcal{C}_{<t}$ be the set of old classes and $\mathcal{C}_t$ be the current new classes. For a current-task sample x, we define the old-new boundary margin as

$$m(x) = \max_{o \in \mathcal{C}_{<t}} s_o(x) - \max_{n \in \mathcal{C}_t} s_n(x). \tag{4}$$

When $m(x) > 0$, at least one old class outranks all current new classes, meaning that x lies on the old

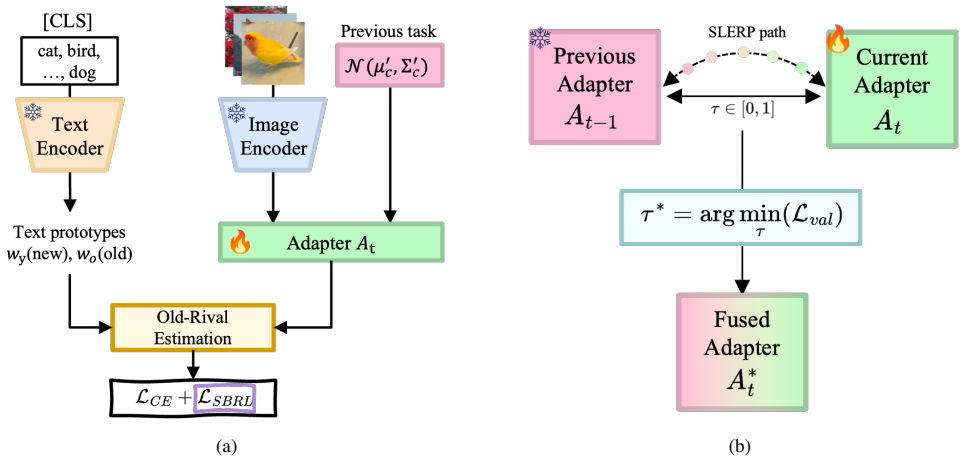


Figure 3: Framework overview. (a) Current-task adaptation with frozen CLIP encoders, CLIP-guided old-rival estimation, and SBRL. (b) Task-end consolidation with MC-SMERG, which merges the previous and current adapters into a single fused adapter for future tasks.

side of the old-vs-new decision boundary. Thus, the fraction of samples with positive margins directly measures old-dominant boundary violations.

For supervised optimization, we use a related label-aware event. For a labeled new-class sample (x, y) with $y \in \mathcal{C}_t$, define

$$B_y(x) = \mathbb{I}\left[\max_{o \in \mathcal{C}_{<t}} s_o(x) > s_y(x)\right]. \quad (5)$$

Equation 4 is a task-level diagnostic: it asks whether the old side outranks the entire current new-class region. Equation 5 is a label-aware training event: it asks whether the true new class is overruled by an old rival. Our loss in Section 3.4 is designed as a smooth surrogate of this label-aware boundary event.

This view also clarifies why new-tail collapse is amplified in frozen-CLIP LTCIL. Consider the old-class logits around a new sample as random variables with marginal CDF F_{old}. Under a simple independent-logit approximation over an effective set of K old rivals, conditioning on the true new-class logit $s_y(x)$ gives

$$\Pr(B_y(x) = 1) \approx 1 - F_{\text{old}}(s_y(x))^K. \quad (6)$$

Here, K denotes the effective number of competitive old rivals rather than the raw number of old classes, since semantically correlated old logits can form local clusters in CLIP space. This approximation is intended as a mechanistic decomposition rather than a complete probabilistic model of correlated CLIP logits; it exposes the main quantities controlling old-dominant boundary risk. Importantly, the SBRL objective introduced in Section 3.4 does not rely on this independence approximation.

Setting	Target→Old	Non-target→Old	Target Acc.	frac_pos
Balanced	0.122	0.127	0.818	0.122
Long-tailed	0.328	0.137	0.577	0.327

Table 1: Controlled evidence for the role of tail scarcity. Only the supervision distribution is changed; class identity, task order, replay, and model capacity are fixed.

First, **semantic overlap** shifts the old-logit distribution upward: because CLIP aligns image and text representations in a shared space, old classes semantically close to a new class can produce competitive logits for new-tail samples. Second, **tail scarcity** weakens the true new-class logit: with few samples, the adapter receives limited supervision to reliably align new-tail images with their text prototype. Third, **old-class expansion** increases the effective number of old rivals as tasks accumulate, making it more likely that some old class dominates. Together, these factors shift Eq. 4 toward positive values, causing new-tail samples to be absorbed into old-class regions.

To isolate tail scarcity, the factor most directly controllable in the data distribution, we conduct a controlled comparison in which class identity, task order, replay, and model capacity are fixed, and only the supervision distribution is changed from balanced to long-tailed. As shown in Table 1, long-tailed supervision selectively increases Target→Old confusion from 0.122 to 0.328 and frac_pos from 0.122 to 0.327, while Non-target→Old confusion remains nearly unchanged. This supports the second factor in our boundary-

risk view: sparse tail supervision selectively amplifies old-dominant boundary collapse. Additional details and full metrics are provided in Appendix A.

This boundary-risk view motivates the rest of our method. Rather than treating new-tail collapse only as a post-hoc diagnostic, we estimate the effective old rivals that drive F_{old} upward and directly suppress their old-side pressure during training. We first describe how these rivals are estimated using CLIP's vision-language geometry, and then introduce a smooth boundary-risk loss to optimize Eq. 5.

3.3 CLIP-Guided Effective Old-Rival Estimation

The boundary-risk view in Section 3.2 suggests that new-tail collapse is driven by a small set of old classes that produce competitive logits for a new sample. Naively considering all old classes is inefficient and can introduce many irrelevant negatives. We therefore estimate an *effective old-rival set* for each new class using the stable vision-language geometry of CLIP.

For a new class $y \in \mathcal{C}_t$, we construct a class-level candidate set from two complementary sources. First, we use CLIP text-space similarity to identify semantically plausible old rivals. Given normalized text prototypes $\{w_c\}$, we retrieve the top-k_{text} old classes most similar to y:

$$\mathcal{N}_{\text{text}}(y) = \text{TopK}_{o \in \mathcal{C}_{<t}} \left(w_o^\top w_y,\ k_{\text{text}} \right). \quad (7)$$

These text-induced rivals capture task-invariant semantic overlap, which remains available even before the current task is fully adapted. This makes the rival estimation language-grounded: CLIP text prototypes provide a stable semantic prior for identifying boundary-critical old-new pairs.

Second, we use adapter-aware image-side evidence from previously learned classes. For each old class o, we store the mean frozen image feature $\bar{f}_o$ and map it through the previous adapter to obtain an adapted old prototype:

$$\hat{\mu}_o = \text{norm}\left(A_{t-1}(\bar{f}_o) \right). \quad (8)$$

We then retrieve old classes whose adapted image prototypes are most aligned with the new text prototype:

$$\mathcal{N}_{\text{img}}(y) = \text{TopK}_{o \in \mathcal{C}_{<t}} \left(\hat{\mu}_o^\top w_y,\ k_{\text{img}} \right). \quad (9)$$

While text similarity captures semantic plausibility, the image-side criterion reflects the current old-class representation geometry induced by the adapter.

The class-level old-rival set is the union of the two sources:

$$\mathcal{O}_{\text{cand}}(y) = \mathcal{N}_{\text{text}}(y) \cup \mathcal{N}_{\text{img}}(y). \quad (10)$$

Figure 4 shows that our method better preserves the CLIP text-space neighborhood structure after continual training, supporting the use of text-guided old-rival candidates.

During training, we refine this class-level set into a sample-level effective rival set by selecting the top-M old rivals according to the current logits:

$$\mathcal{O}_M(x, y) = \text{TopM}_{o \in \mathcal{O}_{\text{cand}}(y)} s_o(x). \quad (11)$$

Figure 5 summarizes how $\mathcal{O}_{\text{cand}}(y)$ is refined into $\mathcal{O}_M(x, y)$ and used by SBRL.

3.4 Smooth Boundary-Risk Loss

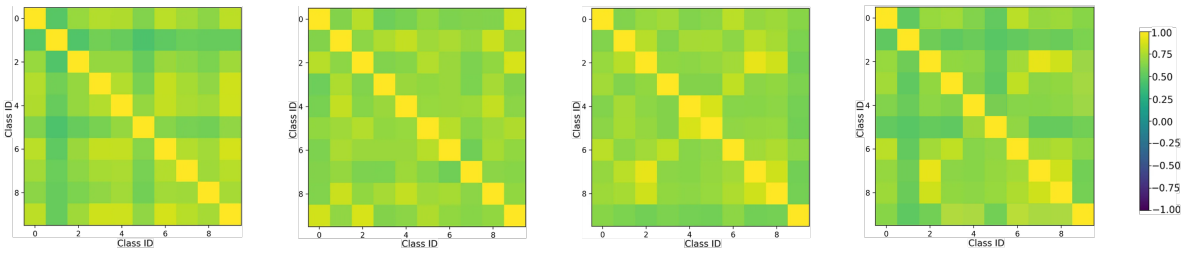
Given the effective old-rival set $\mathcal{O}_M(x, y)$ from Section 3.3, we define a training objective that directly targets old-dominant boundary violations. The label-aware event in Eq. 5 is determined by whether an old rival exceeds the true new-class logit. Rather than treating all old–new pairs uniformly, we first model which pairs repeatedly exhibit boundary violations and then aggregate their risk-adjusted logits into a smooth old-side pressure.

Pairwise risk-adaptive margin. A fixed global margin cannot capture the fact that some old–new pairs repeatedly violate the boundary while others are largely harmless. For each observed old–new pair (o, y), we therefore maintain a pairwise margin $\Delta_{o,y}$ that tracks its recent positive old-over-new overshoot:

$$\Delta_{o,y} \leftarrow (1 - \rho)\Delta_{o,y} + \rho \left[s_o(x) - s_y(x) \right]_+, \quad (12)$$

where $[a]_+ = \max(0, a)$ and ρ is the update rate. We bound $\Delta_{o,y}$ within $[\Delta_{\min}, \Delta_{\max}]$ for stability. Thus, pairs that repeatedly produce large old-over-new violations receive larger margins, whereas non-violating pairs remain weakly constrained. This pairwise memory allows the objective to focus on recurring boundary-critical confusions rather than uniformly separating all old–new pairs.

Smooth multi-rival old-side pressure. A direct implementation of the boundary event could use the largest old-rival logit. However, a hard maximum assigns gradients only to the single largest rival and



(a) Text embedding (b) Continual-CLIP (c) RAPF (d) Ours

Figure 4: Semantic-neighborhood preservation for CLIP-guided old-rival estimation. We visualize class-level similarity structures after the final task. (a) CLIP text-space reference. (b,c) Baselines exhibit distorted neighborhood patterns after continual training. (d) Our method preserves the text-induced structure, supporting the use of text-guided candidates for effective old-rival estimation.

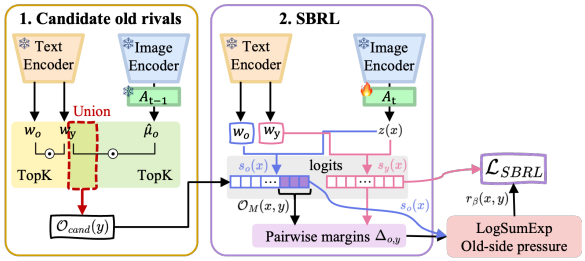


Figure 5: CLIP-guided old-rival estimation and SBRL. We first build a class-level candidate set $\mathcal{O}_{\text{cand}}(y)$ from text-space and adapter-aware image-side rivals, then select top-M sample-level rivals $\mathcal{O}_M(x,y)$. SBRL computes a LogSumExp old-side pressure over these rivals with pairwise adaptive margins and penalizes it when it exceeds the true new-class logit.

can cause unstable rival switching when several old classes are similarly competitive. We therefore define the risk-adjusted old-rival score $u_o(x,y) = s_o(x) + \Delta_{o,y}$ and aggregate the top-M rivals using LogSumExp:

$$r_\beta(x,y) = \frac{1}{\beta} \log \sum_{o \in \mathcal{O}_M(x,y)} \exp(\beta u_o(x,y)), \tag{13}$$

where $\beta > 0$ controls the sharpness of the smooth maximum.

Let $U^*(x,y) = \max_{o \in \mathcal{O}_M(x,y)} u_o(x,y)$ denote the risk-adjusted old-side maximum. Importantly, because $|\mathcal{O}_M(x,y)| = M$, we can bound $r_\beta(x,y)$ as follows:

$$U^*(x,y) \le r_\beta(x,y) \le U^*(x,y) + \frac{\log M}{\beta}. \tag{14}$$

Hence, $r_\beta(x,y)$ serves as a smooth upper approximation to $U^*(x,y)$. As β increases, it approaches this hard maximum, whereas for finite β, gradients are distributed across multiple competitive old rivals. Placing $\Delta_{o,y}$ inside the LogSumExp further ensures that pairs with larger historical overshoot contribute more strongly to the old-side pressure.

Smooth Boundary-Risk Loss. We finally compare the smooth old-side pressure with the true new-class logit:

$$\mathcal{L}_{\text{SBRL}} = \mathbb{E}_{(x,y)\sim\mathcal{B}_{\text{tail}}} \left[r_\beta(x,y) - s_y(x) \right]_+, \tag{15}$$

where $\mathcal{B}_{\text{tail}}$ denotes the current-task tail samples in a mini-batch. The loss is activated only when the risk-adjusted old-side pressure exceeds the true new-class logit. When $\Delta_{o,y} = 0$, it reduces to a smooth multi-rival surrogate of the original old-over-new boundary hinge in Eq. 5; the adaptive margins extend this objective by requiring larger separation for repeatedly boundary-critical pairs.

Compared with standard hard-negative objectives, SBRL therefore makes two boundary-specific design choices: (i) it replaces single-rival hard selection with a smooth multi-rival pressure derived from the old-side maximum in the diagnosed boundary event, and (ii) it uses pairwise adaptive margins to retain recurring old–new overshoot patterns. Section 4.4 empirically isolates both design choices.

Finally, the adapter is trained with the standard cross-entropy loss and SBRL:

$$\mathcal{L}_t = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{SBRL}}, \tag{16}$$

where λ controls the strength of boundary-risk minimization.

3.5 Task-wise Consolidation with MC-SMERG

After training on task t, the updated adapter contains boundary corrections for the newly introduced classes. However, simply replacing the previous adapter with the current one can disturb decision

regions established in earlier tasks, while maintaining a separate adapter per task increases inference cost and requires task identity. We therefore consolidate the previous adapter W_{old} and the current task adapter W_{new} into a single adapter.

We use **M**ode-**C**onnectivity **S**pherical **Merge** (MC-SMERG), a lightweight interpolation-based consolidation strategy. Motivated by mode connectivity observations in neural networks (Draxler et al., 2018; Garipov et al., 2018), we construct a spherical interpolation path between the vectorized adapter weights using SLERP (Shoemake, 1985):

$$W(\tau) = \frac{\sin((1-\tau)\theta)}{\sin\theta} W_{\text{old}} + \frac{\sin(\tau\theta)}{\sin\theta} W_{\text{new}}, \quad \tau \in [0,1], \tag{17}$$

where $\theta = \arccos\left(\frac{\langle \text{vec}(W_{\text{old}}), \text{vec}(W_{\text{new}})\rangle}{\|W_{\text{old}}\|\|W_{\text{new}}\|}\right)$ is the angle between the vectorized adapter weights.

To select the interpolation coefficient, we evaluate a small grid of τ values using a virtual validation set sampled from stored class-wise statistics (μ_c, Σ_c) of all seen classes. We choose

$$\tau^* = \arg\min_{\tau} \mathcal{L}_{\text{CE}}(\mathcal{V}_{\text{virt}}; W(\tau)), \tag{18}$$

and update the adapter as $W_{\text{fused}} = W(\tau^*)$. This selection requires no real exemplars and introduces negligible task-end overhead. Thus, MC-SMERG serves as an efficient consolidation step that preserves a single adapter while retaining the boundary corrections learned across tasks. Appendix D.3 visualizes the surrogate loss along this path, supporting the use of interpolation-based consolidation in our setting.

4 Experiments

4.1 Experimental Setting

Datasets. We evaluate our method on three benchmarks: ImageNetSubset, CIFAR100 (Krizhevsky et al., 2009), and Food101 (Bossard et al., 2014). Following the standard long-tailed CIL protocol (Liu et al., 2022), we construct long-tailed variants ImageNetSubset-LT, CIFAR100-LT, and Food101-LT with imbalance ratio $\rho = 0.01$. For ImageNetSubset-LT, we further evaluate scalability by considering 100-, 200-, and 300-class settings.

Settings. We adopt the class-incremental setup where task identity is unavailable at inference time. Unless otherwise stated, we use the shuffled scenario, which randomizes class order and produces more realistic task-wise imbalance than ordered splits. We report average incremental accuracy (Avg) and final accuracy after the last task (Last). Unless otherwise noted, results use a single default seed and class order; Section 4.5 reports mean $\pm$ standard deviation across multiple seeds and orders. All methods use CLIP ViT-B/16 pre-trained by OpenAI (Radford et al., 2021). Implementation details are provided in Appendix B.

Baselines. We compare against both exemplar-based and exemplar-free CLIP-CIL methods. PROOF (Zhou et al., 2025b) and CLAP4CLIP (Jha et al., 2024) are included as exemplar-based methods. Exemplar-free baselines include Continual-CLIP (Thengane et al., 2022), RAPF (Huang et al., 2024), MoE-Adapters4CL (Yu et al., 2024), and MG-CLIP (Huang et al., 2025).

4.2 Main Results on ImageNetSubset-LT

Table 2 reports results on ImageNetSubset-LT across 100-, 200-, and 300-class settings with both 5-task and 10-task protocols. Our method consistently achieves the best Avg and Last accuracy across all settings while remaining exemplar-free. Compared with the strongest exemplar-free competitors, our method maintains a stable advantage as the number of classes and tasks increases. For example, in the 10-task setting, our method improves over RAPF by 2.16, 1.71, and 2.66 points in Avg accuracy on the 100-, 200-, and 300-class settings, respectively. Our method matches or exceeds exemplar-based baselines without storing real exemplars, and consistently outperforms the recent MG-CLIP across all settings. This suggests that preserving global modality-gap structure alone may be insufficient to resolve local old-new boundary violations under severe long-tailed supervision.

4.3 Fine-Grained Analysis on Tail-Class Performance

To verify whether the gains come from improved tail-class handling, we partition classes into *Low*, *Mid*, and *High* frequency groups according to their training sample counts. Tables 3 and 4 summarize the results on ImageNetSubset-100 and ImageNetSubset-200 under the 10-task setting. Our method achieves the highest accuracy in the data-scarce *Low* group on both benchmarks. On ImageNetSubset-100, it improves the Low-group accuracy over RAPF by 9.14 points (89.43 vs. 80.29). On ImageNetSubset-200, it improves the Low-group accuracy over RAPF by 3.24 points

Methods	Exemplar	100 Classes				200 Classes				300 Classes			
		5		10		5		10		5		10	
		Last(%)	Avg(%)	Last(%)	Avg(%)	Last(%)	Avg(%)	Last(%)	Avg(%)	Last(%)	Avg(%)	Last(%)	Avg(%)
PROOF	✓	86.56	90.20	84.86	89.99	71.78	80.41	72.53	81.61	75.79	81.41	75.35	82.20
CLAP4CLIP	✓	87.25	90.62	87.19	91.43	73.36	79.80	73.92	80.77	76.30	82.52	75.08	83.22
Continual-CLIP	✗	58.60	64.61	58.60	67.69	68.34	75.01	68.34	76.72	72.31	78.41	72.31	79.48
RAPF	✗	86.12	88.00	84.02	89.41	74.92	78.93	74.10	81.15	77.10	80.52	75.84	82.89
MoE-Adapters4CL	✗	87.08	90.68	86.14	90.29	74.66	80.29	73.46	81.51	76.50	82.55	74.62	82.91
MG-CLIP	✗	83.12	87.65	85.78	90.20	72.44	79.59	74.08	81.49	74.34	79.65	75.77	82.14
Ours	✗	**88.78**	**91.55**	**87.38**	**91.57**	**76.80**	**81.85**	**76.41**	**82.86**	**77.96**	**83.89**	**77.42**	**85.55**

Table 2: Comparison on ImageNetSubset-LT benchmarks. '5' and '10' represent the number of tasks, and 'Last' and 'Avg' denote the last and average accuracy across tasks.

Method	Low	Mid	High	Last	Avg
CLAP4CLIP	86.84	86.69	**88.05**	87.19	91.43
RAPF	80.29	87.55	84.48	84.02	89.41
MoE-Adapters4CL	88.29	87.47	83.33	86.14	90.29
Ours	**89.43**	**88.00**	84.77	**87.38**	**91.57**

Table 3: ImageNetSubset-100 (10 tasks).

Method	Low	Mid	High	Last	Avg
CLAP4CLIP	69.00	75.24	**77.73**	73.92	80.77
RAPF	73.90	76.48	72.76	74.10	81.15
MoE-Adapters4CL	71.20	75.13	73.71	73.46	81.51
Ours	**77.14**	**78.33**	73.86	**76.41**	**82.86**

Table 4: ImageNetSubset-200 (10 tasks).

Ablation	Last	Low	$\overline{\text{frac_pos}} \downarrow$
Adapter only	69.84	68.52	0.386
Adapter + SG	72.16	72.03	0.314
Adapter + SG + MC-SMERG	73.78	74.11	0.277
Adapter + SG + SBRL	75.42	77.86	0.156
Full	**76.80**	**79.23**	**0.136**

Table 5: Ablation of framework components (SG, SBRL, MC-SMERG).

(77.14 vs. 73.90).

These Low-group gains align with our boundary-risk view: SBRL targets the regime most vulnerable to new-tail collapse rather than merely lifting average accuracy.

4.4 Ablation and Component Analysis

Tables 5 and 6 disentangle framework components and SBRL-specific design choices on ImageNetSubset-200-LT with 5 tasks. Table 5 shows that SBRL provides a substantial gain independently of MC-SMERG (72.16→75.42 Last), while MC-SMERG provides complementary task-end consolidation; their combination achieves the best Last/Low accuracy and lowest boundary-

Ablation	Last
Baseline + MC-SMERG	73.78
+ single-rival hinge (fixed margin)	74.53
+ multi-rival LogSumExp (fixed margin)	75.79
+ $\mathcal{L}_{\text{SBRL}}$ (multi-rival + adaptive margin)	**76.80**

Table 6: Ablation of SBRL design choices.

Setting	Method	Avg (%)	Last (%)
3 seeds, fixed order	RAPF	79.08 ± 0.29	74.78 ± 0.18
	Ours	$\mathbf{81.95 \pm 0.05}$	$\mathbf{76.73 \pm 0.06}$
5 shuffled orders	RAPF	79.62 ± 0.76	74.80 ± 0.39
	Ours	$\mathbf{82.48 \pm 0.31}$	$\mathbf{76.73 \pm 0.05}$

Table 7: Robustness to random seeds and class-order variation on the shuffled LT-CIL setting. We report mean ± standard deviation of Avg and Last accuracy.

violation rate.

Table 6 progressively isolates the SBRL design. A single-rival boundary hinge improves the baseline from 73.78 to 74.53 Last, multi-rival LogSumExp further reaches 75.79, and the pairwise adaptive margin yields 76.80. Thus, direct boundary correction, multi-rival smoothing, and pair-specific risk adaptation each contribute to the final gain.

Under the same pipeline, conventional imbalance objectives achieve at most 74.46 Last and 75.46 Low accuracy with $\overline{\text{frac_pos}} \geq 0.233$, whereas SBRL obtains 76.80, 79.23, and 0.136, respectively (Appendix E.1).

4.5 Robustness and Mechanism Validation

Robustness to seeds, class orders, and backbone. Table 7 reports robustness on ImageNetSubset-200-LT across three random seeds with a fixed class order and five independently shuffled class orders. Our method consistently outperforms RAPF in both settings, indicating that the gains are not attributable to optimization randomness or a favor-

able class ordering.

The gain also transfers to OpenAI CLIP ViT-B/32 without backbone-specific tuning. Under the same split, class order, and training protocol, our method improves RAPF from 74.89 to 77.11 Last accuracy and from 79.38 to 82.24 Avg accuracy. This supports that both new-tail collapse and the effectiveness of SBRL are not specific to ViT-B/16. Full backbone-specific Low-group and boundary-violation results are reported in Appendix E.2.

Robustness to long-tail severity. We additionally vary the imbalance ratio while keeping the backbone, task protocol, and training setting fixed. Across $\rho \in \{0.1, 0.05, 0.02\}$, our method consistently outperforms both RAPF and MoE-Adapters4CL, with the advantage becoming larger under more severe imbalance. Together with the default $\rho = 0.01$ results, this shows that SBRL is not specialized to a particular long-tail severity. Detailed results are provided in Appendix E.2.

Reliability of old-rival estimation. We next verify whether the proposed class-level candidates actually cover the old classes that dominate new-tail samples. Using an SBRL-free Adapter+SG+MC-SMERG reference model, we define the actual top old violator as the highest-logit old class for each violating new-tail sample. The proposed text∪image candidate set recovers 92.4% of these violators, compared with 82.1% for text-only and 85.7% for image-only candidates.

We further test whether image-side candidates constructed from the previous adapter become stale during current-task training. Refreshing the candidates at every epoch improves Last accuracy only from 76.80 to 76.91, while removing the candidate restriction entirely and searching all old classes reaches 76.96. The small gaps indicate that the static candidate set retains high coverage while avoiding unnecessary old-class comparisons. Full candidate-recall, similarity-bin, and refinement analyses are provided in Appendix E.3.

Cross-modal specificity of new-tail collapse. Finally, we examine whether the observed boundary failure is merely a generic consequence of class imbalance or is amplified by CLIP's cross-modal geometry. Keeping the image encoder, adapter, replay, data stream, and training protocol fixed, we replace the classifier with either a learned visual classifier, image prototypes, or frozen CLIP text prototypes. The class-level Spearman correlation between old-new semantic similarity and boundary-violation rate is substantially stronger with CLIP text prototypes (0.61) than with the learned visual classifier (0.12) or image prototypes (0.18). This supports our interpretation that language-grounded geometry amplifies old-class competition under long-tailed continual adaptation. Full classifier-geometry results are reported in Appendix E.4.

Method	Train Params(↓)	Memory Use(↓)
PROOF	1.3M	18543.60 MiB
CLAP4CLIP	6.8M	1172.76 MiB
RAPF	0.26M	1175.60 MiB
MoE-Adapters4CL	59.8M	18554.60 MiB
MG-CLIP	0.54M	16375.50 MiB
Ours	**0.26M**	**1171.30 MiB**

Table 8: Comparison of training parameters and memory usage during training on CIFAR 100-LT.

4.6 Efficiency and Consolidation

Table 8 compares trainable parameters and training-time memory usage on CIFAR100-LT. Our method uses only 0.26M trainable parameters, matching lightweight adapter-based baselines and using substantially fewer trainable parameters than MoE-Adapters4CL. It also has the lowest reported training memory among the compared methods.

MC-SMERG introduces only a small task-end overhead because it evaluates a lightweight interpolation grid over adapter weights using virtual validation features from stored class statistics. In our implementation, the fusion step takes only 8.3 ms per task on average. Additional interpolation-path analyses and task-end overhead details are provided in Appendix D.

5 Conclusion

We studied frozen-CLIP long-tailed CIL and identified *new-tail collapse*, an old-dominant boundary failure where newly introduced tail classes are absorbed into previously learned regions. Motivated by a boundary-risk analysis linking CLIP semantic overlap, tail scarcity, and old-class expansion, we proposed a CLIP-specific framework with effective old-rival estimation, Smooth Boundary-Risk Loss, and MC-SMERG. Across ImageNetSubset-LT, CIFAR100-LT, and Food101-LT, our method consistently improves over CLIP-based CIL baselines while substantially reducing boundary violations, especially in high-similarity and low-shot regimes.

Limitations

We focus on parameter-efficient frozen-CLIP adaptation, a practical setting for CLIP-based CIL. Future work can examine whether the same boundary-risk dynamics appear under full fine-tuning, prompt-only adaptation, larger vision-language models, or non-classification tasks such as retrieval, detection, segmentation, and multi-modal reasoning. Our old-rival estimation relies on CLIP text prototypes, and may therefore be sensitive to ambiguous class names, prompt templates, or domains where CLIP's text-space neighborhoods do not reflect visual confusability. Finally, the boundary-risk analysis is a simplified mechanistic interpretation rather than a complete theory of CLIP representation geometry.

Ethical Considerations

This work uses CLIP and existing image-classification benchmarks, including ImageNet, CIFAR-100, and Food101, and does not introduce new human-subject data. Our method is exemplar-free and relies on stored class-wise feature statistics rather than raw training images, which can reduce memory requirements and potential privacy risks associated with retaining real exemplars. The lightweight adapter-based design also reduces the computational burden compared with heavier continual adaptation methods.

At the same time, our method inherits limitations of the underlying CLIP model. Because CLIP is trained on large-scale web data, its visual and textual representations may encode social, cultural, or geographic biases. Improving tail-class performance does not by itself guarantee fair or safe behavior, especially when tail categories correspond to underrepresented or sensitive groups. Therefore, deployment in high-stakes settings should require additional bias, robustness, and domain-specific safety evaluations. Our experiments are intended to study long-tailed continual learning behavior on standard benchmarks, not to validate the system for safety-critical applications.

Acknowledgment

This research was supported by the AI Seoul Tech Research Support Program of the Seoul Future Foundation. This work was also supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)].

References

Reza Akbarian Bafghi and Danna Gurari. 2023. A new dataset based on images taken by blind people for testing the robustness of image classification models trained for imagenet categories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16261–16270.

S Divakar Bhat, Amit Popat More, Mudit Soni, and Bhuvan Aggarwal. 2026. Adaprior: Bayesian-inspired adaptive prior correction for long-tailed continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10840–10850.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, pages 446–461. Springer.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32.

Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. 2018. End-to-end incremental learning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 233–248.

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

Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. 2018. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pages 1309–1318. PMLR.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. 2018. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31.

Yanan Gu, Muli Yang, Xu Yang, Kun Wei, Hongyuan Zhu, Gabriel James Goenawan, and Cheng Deng. 2025. Dynamic adapter tuning for long-tailed class-incremental learning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8176–8185. IEEE.

Tyler L Hayes and Christopher Kanan. 2020. Lifelong machine learning with deep streaming linear discriminant analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 220–221.

Jiangpeng He. 2024. Gradient reweighting: Towards imbalanced class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16668–16677.

Chenxing Hong, Yan Jin, Zhiqi Kang, Yizhou Chen, Mengke Li, Yang Lu, and Hanzi Wang. 2024. Dynamically anchored prompting for task-imbalanced continual learning. *arXiv preprint arXiv:2404.14721*.

Yijie Hu, Guanyu Yang, Zhaorui Tan, Xiaowei Huang, Kaizhu Huang, and Qiu-Feng Wang. 2025. Covariance-based space regularization for few-shot class incremental learning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 9566–9576. IEEE.

Linlan Huang, Xusheng Cao, Haori Lu, and Xialei Liu. 2024. Class-incremental learning with clip: Adaptive representation adjustment and parameter fusion. In *European Conference on Computer Vision*, pages 214–231. Springer.

Linlan Huang, Xusheng Cao, Haori Lu, Yifan Meng, Fei Yang, and Xialei Liu. 2025. Mind the gap: Preserving and compensating for the modality gap in clip-based continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3777–3786.

Saurav Jha, Dong Gong, and Lina Yao. 2024. Clap4clip: Continual learning with probabilistic finetuning for vision-language models. *arXiv preprint arXiv:2403.19137*.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2020. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*.

Minsoo Kang, Jaeyoo Park, and Bohyung Han. 2022. Class-incremental learning by knowledge distillation with adaptive feature consolidation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16071–16080.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Alex Krizhevsky, Geoffrey Hinton, and 1 others. 2009. Learning multiple layers of features from tiny images.

Xinyi Lai, Luojun Lin, Weijie Chen, and Yuanlong Yu. 2025. A tiny change, a giant leap: Long-tailed class-incremental learning via geometric prototype alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1444–1453.

Zhizhong Li and Derek Hoiem. 2016. Learning without forgetting. In *European conference on computer vision*, pages 614–629. Springer.

Xialei Liu, Yu-Song Hu, Xu-Sheng Cao, Andrew D Bagdanov, Ke Li, and Ming-Ming Cheng. 2022. Long-tailed class incremental learning. In *European Conference on Computer Vision*, pages 495–512. Springer.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. 2019. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2537–2546.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. 2020. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Dilan Gorur, Razvan Pascanu, and Hassan Ghasemzadeh. 2021. Linear mode connectivity in multitask and continual learning. In *International Conference on Learning Representations*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2001–2010.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and 1 others. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. 2017. Continual learning with deep generative replay. In *Advances in neural information processing systems*, volume 30.

Ken Shoemake. 1985. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pages 245–254.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. 2016. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769.

Yu-Ming Tang, Yi-Xing Peng, and Wei-Shi Zheng. 2023. When prompt-based incremental learning does not meet strong pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1706–1716.

Vishal Thengane, Salman Khan, Munawar Hayat, and Fahad Khan. 2022. Clip model is an efficient continual learner. *arXiv preprint arXiv:2210.03114*.

Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. 2019. Large scale incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 374–382.

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. 2024. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23219–23230.

Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, and Yunchao Wei. 2023. Slca: Slow learner with classifier alignment for continual learning on a pre-trained model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19148–19158.

Jiarui Zhao, Libo Huang, Xiangqi Li, Zhulin An, Chuanguang Yang, Yu Wang, Boyu Diao, and Yongjun Xu. 2026. Representation-steered incremental adapter-tuning for class-incremental learning with pre-trained models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18010–18020.

Da-Wei Zhou, Kai-Wen Li, Jingyi Ning, Han-Jia Ye, Lijun Zhang, and De-Chuan Zhan. 2025a. External knowledge injection for clip-based class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3314–3325.

Da-Wei Zhou, Yuanhan Zhang, Yan Wang, Jingyi Ning, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. 2025b. Learning without forgetting for vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Appendix

A Additional Boundary-Risk Analysis and Validation

We provide additional details for the boundary-risk view introduced in Section 3.2. The main paper defines the task-level margin $m(x)$ and the label-aware collapse event $B_y(x)$. Here, we derive the approximation used in Eq. 6, discuss its assumptions, and empirically validate its predictions.

A.1 Derivation of the Boundary-Risk Approximation

Let F_{old} denote the marginal CDF of old-class logits around a new sample (x, y). Conditioning on $s_y(x)$ and using an independent-logit approximation over an effective set of K old rivals, we have

$$\Pr\left(\max_{o\in\mathcal{C}_{<t}} s_o(x) \le s_y(x)\right) \approx F_{\text{old}}(s_y(x))^K. \tag{A.19}$$

Therefore,

$$\Pr(B_y(x)=1) \approx 1 - F_{\text{old}}(s_y(x))^K. \tag{A.20}$$

This approximation is used only to expose qualitative dependencies. It shows that old-dominant collapse becomes more likely when the old-logit distribution shifts upward, when the true new-class logit decreases, or when the effective number of competitive old rivals increases.

For analytical illustration, if $s_o(x) \sim \mathcal{N}(\mu_{\text{old}}, \sigma^2_{\text{old}})$, then

$$\Pr(B_y(x)=1) \approx 1 - \Phi\left(\frac{s_y(x)-\mu_{\text{old}}}{\sigma_{\text{old}}}\right)^K. \tag{A.21}$$

We do not assume that CLIP logits are exactly Gaussian. Cosine logits are bounded and may be multi-modal due to semantic clustering. The Gaussian form is only an illustrative instantiation of the general CDF-based expression.

A.2 Assumptions and Interpretation

The independent-logit assumption is also a simplification. In practice, old-class logits are correlated because CLIP organizes semantically related classes into local neighborhoods. Thus, K should be interpreted as an effective number of competitive old rivals rather than the raw number of old classes. Semantic clustering can reduce the effective independent count, but it can also increase the local old-rival logit mean for semantically close classes.

The three factors in Section 3.2 correspond to the terms in the approximation. Semantic overlap shifts F_{old} toward larger logits. Tail scarcity weakens or destabilizes $s_y(x)$ because the adapter receives limited supervision for the new class. Old-class expansion increases the old-side maximum through a max-over-classes effect. For example, under the Gaussian illustration, the expected maximum old logit grows approximately as

$$\mathbb{E}\left[\max_o s_o(x)\right] \approx \mu_{\text{old}} + \sigma_{\text{old}}\sqrt{2\log K}, \tag{A.22}$$

which highlights why an expanding old-class pool can become increasingly competitive for new-tail samples.

A.3 Empirical Validation of Boundary-Risk Predictions

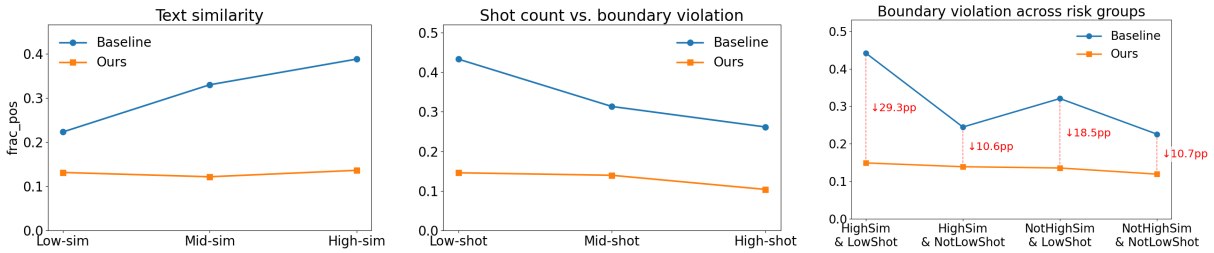
The boundary-risk view leads to the following predictions: (i) violations should increase with old-new text similarity, (ii) violations should increase as the number of training samples decreases, and (iii) boundary-risk minimization should reduce violations most strongly when high similarity and low shot count coincide.

We evaluate these predictions by grouping current-task new classes according to old-new text similarity and training-shot count, and measuring class-balanced $\mathrm{frac_pos}$ over their test samples. Here, $\mathrm{frac_pos}$ denotes the fraction of samples satisfying

$$\max_{o\in\mathcal{C}_{<t}} s_o(x) - \max_{n\in\mathcal{C}_t} s_n(x) > 0. \tag{A.23}$$

Text similarity bins are computed using the top-k average similarity between each new-class text prototype and old-class text prototypes. Shot-count bins are computed from the number of training samples in the current task.

Figure 6 validates these predictions. Under the baseline, boundary violations increase with old-new text similarity and are concentrated in low-shot classes. The largest reduction appears in the high-risk group where high semantic similarity and low shot count coincide. These results support our interpretation that new-tail collapse arises from the interaction of semantic overlap, tail scarcity, and old-class competition, and that SBRL suppresses the corresponding old-dominant boundary violations.



(a) (b) (c)

Figure 6: Empirical validation of the boundary-risk mechanism on ImageNetSubset-200-LT. (a) Boundary violations increase with old-new text similarity under the baseline, while our method reduces and flattens this dependency. (b) Violations are concentrated in low-shot classes, supporting that tail scarcity increases old-dominant boundary failures. (c) The largest reduction appears in the predicted high-risk group, where high old-new semantic similarity and low shot count coincide. All panels report class-balanced frac_pos averaged within each current-task new class. Red annotations denote absolute reductions in percentage points.

B Training Details

We implement all methods in PyTorch 2.5.1 and run experiments on a single NVIDIA RTX 4070 Ti Super GPU. We follow the standard CLIP-based CIL setup (Huang et al., 2024): the CLIP image/text encoders are frozen and only a lightweight adapter is trained. The adapter is optimized with cross-entropy on current-task data and statistic-based feature replay, where Gaussian pseudo-features are sampled from stored class-wise statistics of previously seen classes.

For our method, we set the SBRL loss weight to $\lambda = 1.0$, the number of sample-level old rivals to $M = 3$, and the LogSumExp sharpness to $\beta = 5$. For CLIP-guided old-rival estimation, we use $k_{\text{text}} = k_{\text{img}} = 5$. Pairwise risk-adaptive margins are initialized to 0.1 and updated online with an exponential moving average of positive old-over-new overshoot. For MC-SMERG, we use a grid search with 9 interpolation steps per task to select the merge point. All other hyperparameters follow (Huang et al., 2024) unless specified otherwise.

B.1 Hyperparameter Sensitivity

We analyze the sensitivity of the main SBRL hyperparameters on ImageNetSubset-200-LT with 5 tasks. As shown in Table 9, performance remains stable across reasonable ranges of M, β, $k_{\text{text}} = k_{\text{img}}$, and λ. Across all tested settings, Last accuracy varies within 0.5 points, and the boundary-violation rate frac_pos remains consistently low (0.13–0.17), indicating that boundary-risk reduction is not sharply peaked at the default configuration.

We also evaluate the sensitivity of MC-SMERG to the number of interpolation search steps. Figure 7 shows that performance is stable across a reasonable range of search steps and peaks around our default configuration. This indicates that the merge selection is not highly sensitive to the exact grid resolution.

Hyperparameter	Value	Last (%)	Avg (%)	frac_pos
M	1	76.31	81.47	0.173
	3	76.80	81.85	0.136
	5	76.64	81.76	0.142
β	2	76.52	81.68	0.151
	5	76.80	81.85	0.136
	10	76.58	81.70	0.146
$k_{\text{text}} = k_{\text{img}}$	3	76.48	81.61	0.149
	5	76.80	81.85	0.136
	10	76.70	81.78	0.138
λ	0.5	76.35	81.54	0.162
	1.0	76.80	81.85	0.136
	2.0	76.57	81.65	0.131

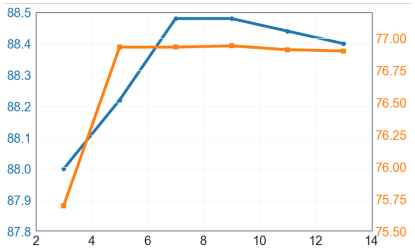
Table 9: Hyperparameter sensitivity of SBRL. Default values are $M = 3$, $\beta = 5$, $k_{\text{text}} = k_{\text{img}} = 5$, and $\lambda = 1.0$.

B.2 Dataset statistics.

Dataset	Train	Test
ImageNetSubset-200-LT	56,171	10,000
CIFAR100-LT	10,847	10,000
Food101-LT	16,454	25,250

Table 10: Dataset statistics under the long-tailed setting. Train denotes the long-tailed training split; the standard test split is retained for evaluation.

Table 10 summarizes the number of training and and test samples used in our long-tailed bench-



(a) Hyperparameter τ

Figure 7: Hyperparameter Sensitivity Analysis. Performance by search steps(τ) for MC-SMERG.

marks. We construct long-tailed training sets while retaining the corresponding test sets for evaluation. No separate validation split is used. The 5-task and 10-task protocols use the same underlying training and test sets and differ only in how classes are partitioned across tasks.

Dataset	Task	Train Samples
ImageNet-200	1	8,497
	2	10,724
	3	14,271
	4	10,068
	5	12,611
CIFAR-100	1	2,025
	2	1,663
	3	2,688
	4	2,346
	5	2,125
Food-101	1	2,322
	2	2,957
	3	2,592
	4	4,381
	5	4,202

Table 11: Task-wise training-set sizes under the 5-task shuffled setting. ImageNetSubset-200 and CIFAR100 contain 40 and 20 new classes per task, respectively; Food101 contains 21 classes in the first task and 20 thereafter.

Table 11 further reports the task-wise training-set sizes under the 5-task shuffled setting. ImageNetSubset-200 and CIFAR100 introduce 40 and 20 classes per task, respectively, while Food101 introduces 21 classes in the first task and 20 thereafter. Evaluation is cumulative over all classes observed up to each task.

C Additional Benchmark Results

C.1 Results on CIFAR100-LT

We additionally evaluate our method on CIFAR100-LT with imbalance factor $\rho = 0.01$. As reported in Table 12, our method achieves the best performance in both the 5-task and 10-task settings. In the 5-task setting, our method obtains 87.44 Avg and 81.09 Last accuracy, outperforming the strongest exemplar-free baseline MoE-Adapters4CL. In the 10-task setting, our method reaches 89.52 Avg and 83.22 Last accuracy, slightly surpassing the recent MG-CLIP baseline. These results show that our boundary-risk minimization strategy transfers beyond ImageNetSubset-LT to a standard long-tailed benchmark.

Methods	Exemplar	CIFAR100-LT			
		5		10	
		Last(%)	Avg(%)	Last(%)	Avg(%)
PROOF	✓	72.88	79.71	72.26	80.68
CLAP4CLIP	✓	73.00	78.83	72.85	71.77
Continual-CLIP	✗	68.93	75.72	68.93	77.28
RAPF	✗	78.60	85.58	75.79	83.08
MoE-Adapters4CL	✗	80.06	87.24	78.60	86.46
MG-CLIP	✗	80.09	87.08	83.20	89.26
Ours	✗	**81.09**	**87.44**	**83.22**	**89.52**

Table 12: Comparison on CIFAR100-LT benchmarks with imbalance factor $\rho = 0.01$. '5' and '10' represent the number of tasks, and 'Last' and 'Avg' denote the last and average accuracy across tasks.

Tail-group accuracy on CIFAR100-LT. To assess whether the gains are driven by improved tail recognition, we report per-group accuracy on CIFAR100-LT under the 10-task setting. As shown in Table 13, our method achieves the best accuracy on the data-scarce Low group and the best final accuracy. This supports our claim that SBRL improves new-tail learning rather than only increasing aggregate performance.

Method	Exemplar	Low	Mid	High	Last	Avg
CLAP4CLIP	✓	62.27	76.21	**79.97**	72.85	71.77
RAPF	✗	75.62	82.47	75.29	75.79	83.08
MoE-Adapters4CL	✗	79.22	**82.04**	74.29	78.60	86.46
Ours	✗	**80.01**	**82.04**	75.62	**79.22**	**86.52**

Table 13: Per-group accuracy comparison on CIFAR-100 (10 tasks).

C.2 Results on Food101-LT

We further evaluate our method on Food101-LT with imbalance factor $\rho = 0.01$. Since Food101 contains 101 classes, we follow the standard split except that the first task contains one additional class to match the total class count. Table 14 reports results under both 5-task and 10-task protocols.

Our method achieves the best overall performance among all compared approaches. In the 5-task setting, it obtains 91.40 Avg and 87.84 Last

accuracy, outperforming RAPF, MG-CLIP, and exemplar-based baselines. In the 10-task setting, our method reaches 91.51 Avg and 87.95 Last accuracy, again outperforming the strongest compared methods. These results indicate that the proposed boundary-risk minimization framework generalizes to a fine-grained long-tailed benchmark.

Methods	Exemplar	Food101-LT			
		5		10	
		Last(%)	Avg(%)	Last(%)	Avg(%)
PROOF	✓	85.28	88.07	86.10	90.05
CLAP4CLIP	✓	86.31	90.11	87.24	91.10
Continual-CLIP	✗	68.93	75.72	68.93	77.28
RAPF	✗	87.04	89.26	87.38	91.21
MoE-Adapters4CL	✗	84.31	87.06	86.64	88.38
MG-CLIP	✗	83.38	89.02	87.52	91.03
Ours	✗	**87.84**	**91.40**	**87.95**	**91.51**

Table 14: Comparison on Food101-LT benchmarks with imbalance factor $\rho = 0.01$. '5' and '10' represent the number of tasks, and 'Last' and 'Avg' denote the last and average accuracy across tasks.

D MC-SMERG Overhead and Loss Geometry

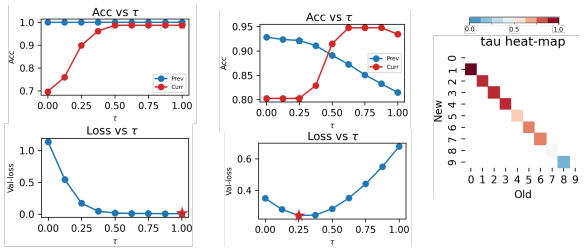
D.1 Task-End Overhead

Table 15 reports the task-end overhead incurred by MC-SMERG. We decompose the overhead into virtual validation set (VVS) construction time and path evaluation time. The total selection overhead is the sum of these two components.

Across tasks, the average task-end overhead is only 0.085 seconds, corresponding to 0.04% of the corresponding task training time. This confirms that MC-SMERG introduces negligible wall-clock overhead while maintaining a single consolidated adapter.

D.2 Interpolation Dynamics

Figure 8 visualizes the interpolation behavior of MC-SMERG in an early task and a later task. The accuracy curves illustrate the stability-plasticity trade-off along the interpolation path. In early tasks, the selected interpolation coefficient favors plasticity toward the current task. In later tasks, the selected coefficient shifts toward stability, reflecting the growing need to preserve previously learned decision regions. The heatmap summarizes this adaptive trend across all tasks.



(a) First task (b) Last task (c) Heat map

Figure 8: Analysis of MC-SMERG dynamics in an early (Task 1) and late (Task 9) step. The accuracy plots show the stability (blue) vs. plasticity (red) trade-off. The validation loss (bottom) is minimized at an optimal τ^* (star), which shifts from a plasticity to a stability focus over time. The heatmap (right) shows this adaptive trend across all tasks.

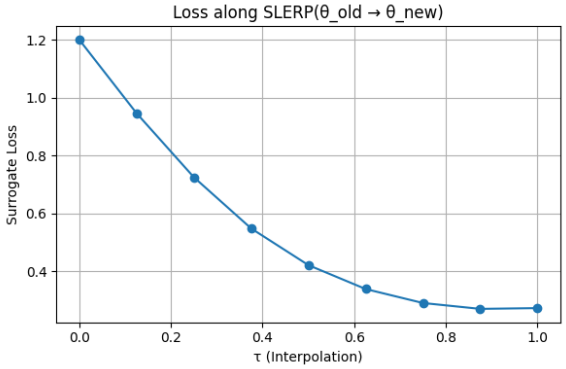


Figure 9: Surrogate loss along the interpolation path $\theta(\tau)$.

D.3 Loss Geometry Along the Interpolation Path

To further understand the effectiveness of MC-SMERG, we analyze the surrogate loss landscape between the old and new adapters. Following mode-connectivity studies (Garipov et al., 2018; Mirzadeh et al., 2021), we evaluate the surrogate loss along the SLERP interpolation path:

$$\theta(\tau) = \mathrm{SLERP}(\theta_{\mathrm{old}}, \theta_{\mathrm{new}}, \tau), \quad \tau \in [0, 1]. \tag{A.24}$$

At each interpolation point, we evaluate the cross-entropy loss on a virtual validation set composed of Gaussian samples generated from stored class-wise statistics.

Figure 9 shows that the surrogate loss changes smoothly along the interpolation path. This supports the use of interpolation-based adapter merging and explains why a lightweight grid search can identify a suitable fused adapter without storing real exemplars.

Task t	Selection overhead (s/task)	VVS construction (s/task)	Path eval (s/task)	Overhead / train (%)
1	0.062	0.035	0.027	0.04
2	0.045	0.036	0.009	0.02
3	0.057	0.050	0.007	0.03
4	0.068	0.061	0.007	0.02
5	0.084	0.074	0.010	0.03
6	0.094	0.087	0.007	0.05
7	0.105	0.097	0.008	0.04
8	0.128	0.119	0.009	0.05
9	0.127	0.119	0.008	0.03
Avg.	0.085	0.075	0.010	0.04

Table 15: **Task-end overhead of MC-SMERG selection.** Selection overhead decomposes into virtual validation set (VVS) construction and low-loss path evaluation. We also report the overhead as a fraction of the corresponding per-task training time.

E Additional Ablations and Mechanism Analyses

Unless otherwise noted, all experiments in this section use ImageNetSubset-200-LT with 5 tasks, OpenAI CLIP ViT-B/16, and the same training protocol and hyperparameters as the main experiments. We provide additional analyses of alternative imbalance objectives, robustness to backbone and long-tail severity, old-rival estimation, and classifier geometry.

E.1 Comparison with Conventional Imbalance Objectives

We examine whether the gains of SBRL can be reproduced by generic hard-example mining or conventional long-tailed classification objectives. We replace SBRL with OHEM (Shrivastava et al., 2016), LDAM (Cao et al., 2019), logit adjustment (Menon et al., 2020), and class-balanced loss (Cui et al., 2019) while keeping the Adapter+SG+MC-SMERG pipeline unchanged. Hyperparameters for all objectives are selected using the same validation protocol.

Objective	Last	Low	$\overline{\text{frac_pos}}\downarrow$
CE	73.78	71.11	0.277
OHEM	73.42	74.15	0.247
LDAM	71.66	73.08	0.243
Logit adjustment	74.46	74.57	0.233
Class-balanced loss	73.91	75.46	0.242
SBRL	**76.80**	**79.23**	**0.136**

Table 16: Comparison with conventional imbalance objectives under the identical Adapter+SG+MC-SMERG pipeline. SBRL achieves the highest Last and Low-group accuracy while most strongly reducing old-dominant boundary violations.

While conventional imbalance objectives improve some tail-class metrics, none directly models the directional old-over-new boundary failure diagnosed in our setting. In particular, SBRL improves Last accuracy by 2.34 points over the strongest alternative, logit adjustment, and reduces $\overline{\text{frac_pos}}$ from 0.233 to 0.136. These results indicate that generic hard mining or global imbalance correction does not fully account for the local old-new competition targeted by SBRL.

E.2 Backbone and Imbalance-Ratio Robustness

Backbone generality. We evaluate OpenAI CLIP ViT-B/32 using the same data split, class order, training protocol, and SBRL hyperparameters as ViT-B/16, without backbone-specific tuning.

Method	Backbone	Last	Avg	Low	$\overline{\text{frac_pos}}\downarrow$
RAPF	ViT-B/16	74.92	78.93	74.69	0.318
RAPF	ViT-B/32	74.89	79.38	74.31	0.334
Ours	ViT-B/16	**76.80**	**81.85**	**79.23**	**0.136**
Ours	ViT-B/32	**77.11**	**82.24**	**78.96**	**0.150**

Table 17: Backbone robustness on ImageNetSubset-200-LT with 5 tasks. ViT-B/32 uses the same SBRL hyperparameters as ViT-B/16 without backbone-specific tuning.

On ViT-B/32, our method improves RAPF by 2.22 Last, 2.86 Avg, and 4.65 Low accuracy points while reducing $\overline{\text{frac_pos}}$ from 0.334 to 0.150. The consistent improvement across both backbones suggests that the observed boundary failure and the effectiveness of SBRL are not specific to ViT-B/16.

Long-tail severity. We further vary the imbalance ratio $\rho \in \{0.1, 0.05, 0.02\}$ while keeping the backbone, task protocol, and all other settings fixed.

Method	$\rho = 0.1$	$\rho = 0.05$	$\rho = 0.02$
RAPF	77.68 / 82.38	75.68 / 80.09	74.98 / 79.18
MoE-Adapters4CL	77.34 / 83.89	76.48 / 83.00	75.26 / 81.64
Ours	**78.43 / 84.04**	**77.69 / 83.48**	**77.34 / 82.66**

Table 18: Robustness to long-tail severity. Each entry reports Last / Avg accuracy on ImageNetSubset-200-LT with 5 tasks.

Together with the default $\rho = 0.01$ setting in the main experiments, our method consistently outperforms both baselines across all tested imbalance ratios. The gains generally become more pronounced as the imbalance becomes more severe, indicating that SBRL is not specialized to a single long-tail severity.

E.3 Old-Rival Estimation and Sample-Level Refinement

We evaluate the reliability of the two-stage old-rival estimation introduced in Section 3.3: class-level candidate construction followed by sample-level current-logit refinement.

Evaluation protocol. We use an SBRL-free Adapter+SG+MC-SMERG model as a reference, with actual violators measured using its task-end fused adapter. For each violating new-tail sample (x, y), we define its actual highest-logit old violator as

$$o^*(x) = \arg\max_{o \in \mathcal{C}_{<t}} s_o^{\text{ref}}(x). \tag{A.25}$$

We report $R@1_{\text{viol}}$, the class-balanced recall that the class-level candidate set contains $o^*(x)$. Recall is first computed within each new-tail class and then averaged equally across classes and tasks; classes with no violations are excluded. Ground-truth test labels are used only for this post-hoc analysis and are never used for candidate construction or training.

Candidate construction. For fair comparison, random, text-only, and image-only restricted candidate sets are size-matched to the proposed text∪image set, whose average size is 8.4 classes. Random recall is averaged over 50 draws.

The union improves $R@1_{\text{viol}}$ to 92.4%, compared with 82.1% for text-only and 85.7% for image-only candidates, confirming their complementarity. The benefit of image-side evidence is especially pronounced for classes with weaker text-space similarity: in Low/Mid/High text-similarity bins, text-only recall is 64.8/82.7/94.1%, whereas the union reaches 87.8/92.1/96.7%.

We also examine potential staleness from constructing image-side candidates using the previous fused adapter A_{t-1}. Refreshing candidates every epoch increases recall only from 92.4% to 93.2% and Last accuracy from 76.80 to 76.91. Using all old classes gives 100% coverage but improves Last accuracy only to 76.96. These small gaps suggest that the static candidate set captures most boundary-critical rivals without requiring repeated candidate updates or comparisons against the entire old-class pool.

Sample-level refinement. We next fix the same text∪image candidate set and vary only how the final rivals are selected. Since SBRL uses $M = 3$ rivals, we measure Overlap@3 with the reference model's actual top-three old-logit rivals. Static top-3 ranks candidates using their class-level normalized text/image similarity, whereas current-logit top-3 uses the sample-specific logits $s_o(x)$.

Current-logit selection achieves 89.1% Overlap@3 with the actual top-three old rivals, compared with 71.8% for static class-level selection. It also gives the highest accuracy and lowest violation rate. Together, these results support the two-stage design: text and image geometry provide high-coverage class-level candidates, while current logits refine them into sample-specific effective rivals.

E.4 Classifier-Geometry Analysis

We examine whether new-tail collapse is solely a generic consequence of long-tailed continual learning or is amplified by CLIP's cross-modal classifier geometry. We keep the frozen image encoder, adapter, statistic-based replay, data stream, and training protocol fixed, while varying only the classifier geometry: (i) a normalized learned visual classifier trained with current data and SG pseudo-features, (ii) normalized image prototypes constructed from stored image-feature statistics, and (iii) frozen CLIP text prototypes.

For each new class, we compute its mean similarity to its top-5 old CLIP text prototypes and measure the Spearman correlation with its boundary-violation rate, averaged across tasks. The substantially larger correlation for CLIP text prototypes (0.61 versus 0.12–0.18 for the learned visual classifier and image prototypes) indicates that semantic proximity in the cross-modal text space is strongly

Candidate strategy	$R@1_{\text{viol}}$	Last	Avg
Random, size-matched	11.3 ± 0.4	74.22 ± 0.18	79.80 ± 0.16
Text-only	82.1	75.54	80.76
Image-only, static A_{t-1}	85.7	75.88	81.05
Text∪Image, static A_{t-1}	**92.4**	**76.80**	**81.85**
Text∪Image, epoch-refreshed	93.2	76.91	81.94
All-old rivals	100.0	76.96	82.00

Table 19: Coverage and performance of different old-rival candidate strategies. Restricted candidate sets are size-matched. The static proposed candidate set captures most actual top old violators while remaining substantially smaller than the full old-class set.

Rival selection	Last	Avg	$\overline{\text{frac_pos}} \downarrow$
All candidates in LogSumExp	75.92	81.10	0.165
Static class-level top-3	76.18	81.31	0.152
Current-logit top-3	**76.80**	**81.85**	**0.136**

Table 20: Effect of sample-level old-rival refinement with the same text∪image class-level candidate set.

coupled with old-dominant violations. SBRL reduces both the violation rate and this correlation, supporting our interpretation that cross-modal geometry amplifies new-tail collapse and that SBRL specifically mitigates this effect.

E.5 Boundary-Critical Confusion Case Study

We further isolate SBRL's effect in a controlled high-risk semantic confusion scenario. We train on two old head classes (*Coffee mug* and *Pot*) and subsequently introduce two semantically close new tail classes (*Tea pot* and *Cup*), selected using CLIP text-space similarity.

Without SBRL, many samples from the semantically related new-tail classes are absorbed into the old classes. SBRL substantially increases correct predictions for *Tea pot* (23→31) and *Cup* (19→40), while performance on the old head classes remains essentially unchanged. This provides a direct instance-level illustration that SBRL improves new-tail decision regions without simply trading off old-class performance.

F Classes of ImageNetSubset

For our ImageNet100, ImageNet200, and ImageNet300 benchmarks, we selected subsets of categories from the full ImageNet1k dataset.

Classes of ImageNet100 To ensure fair comparisons across the 100-class subset, we strictly adhered to the class list and long-tail distribution settings proposed by prior research (Liu et al., 2022).

Classes of ImageNet200 The selection of classes for the ImageNet200 subset was inspired by the class distribution observed in the VizWiz-Classification dataset (Bafghi and Gurari, 2023). The VizWiz dataset primarily contains images captured by blind and low-vision users, often featuring object classes distinct from those typically highlighted in standard benchmarks. By referencing this unique distribution, we aimed to create a challenging and diverse benchmark that effectively evaluates the scalability and robustness of class incremental learning (CIL) methods. The specific class names included in our ImageNet200 subset are listed below:

tiger, ice bear, hog, lion, umbrella, soccer ball, grand piano, laptop, strawberry, shopping cart, racer, sports car, pickup, trailer truck, bookcase, china cabinet, chiffonier, table lamp, file, folding chair, rocking chair, studio couch, toilet seat, desk, wardrobe, Granny Smith apple, orange, lemon, pineapple, banana, pomegranate, corn, upright, chime, acoustic guitar, electric guitar, valley, can opener, screwdriver, loudspeaker, microphone, mouse, electric fan, strainer, space heater, stove, rule, scale, analog clock, digital clock, wall clock, digital watch, projector, sunglasses, computer keyboard, lighter, desktop computer, printer, vending machine, joystick, hook, car wheel, swing, radiator, car mirror, remote control, buckle, seat belt, candle, fly, bee, iron, espresso maker, microwave, toaster, vacuum, dishwasher, refrigerator, washer, Crock Pot, frying pan, wok, teapot, patio, library, restaurant, grocery store, fountain, chainlink fence, picket fence, sliding door, pedestal, mashed potato, bell pepper, broccoli, cauliflower, cucumber, artichoke, mushroom, shower curtain, jean, carton, handkerchief, sandal, ashcan, plate, necklace, pajama, running shoe, chest, modem, tub, tray, book jacket, beer bottle, tile roof, teddy, pop bottle, suit, red wine, Christmas stocking, menu, stage, baseball, sunscreen, lampshade, bow tie, water jug,

Classifier geometry	Last	Low	$\overline{\text{frac_pos}}\downarrow$	Sim.–viol. corr.
Learned visual classifier	73.92	73.44	0.214	0.12
Image prototypes	74.31	74.06	0.231	0.18
CLIP text prototypes	73.78	71.11	0.277	0.61
CLIP text prototypes + SBRL	**76.80**	**79.23**	**0.136**	**0.16**

Table 21: Effect of classifier geometry on new-tail collapse. The first three variants use cross-entropy training; the last adds SBRL. Sim.–viol. corr. denotes the class-level Spearman correlation between old-new semantic similarity and boundary-violation rate.

Task	Class	w/o SBRL	w/ SBRL
Old	Coffee mug	36	36
	Pot	47	48
New	Tea pot	23	31
	Cup	19	40

Table 22: Number of correct predictions in a boundary-critical old-new confusion scenario with and without SBRL.

bucket, soup bowl, chain, mixing bowl, wine bottle, binder, cardigan, sweatshirt, pot, hamper, backpack, cup, radio, dough, lipstick, monitor, vase, paper towel, envelope, perfume, bathtub, hotdog, pillow, toilet paper, cassette, lotion, hair spray, pill bottle, window shade, barrel, washbasin, ballpoint pen, basketball, bath towel, window screen, cellular telephone, mailbox, fire screen, packet, pole, wallet, cassette player, comic book, piggy bank, street sign, Windsor tie, purse, television, measuring cup, espresso, pizza, wooden spoon, saltshaker, ball player, goblet, water bottle, soap dispenser, T-shirt, plastic bag, diaper, doormat, Loafer, ice cream, pretzel, quilt, tape player, iPod, pitcher, sock, CD player, cheeseburger, coffee mug.

Classes of ImageNet300 For the 300 class subset, we combined classes from the 100-class and 200-class subsets. To resolve the overlap of 23 duplicated classes between these subsets, we randomly selected alternative classes to replace them, maintaining a distinct and representative set. The specific class names are listed below:

tench, goldfish, great white shark, tiger shark, hammerhead shark, electric ray, stingray, hen, ostrich, brambling, house finch, junco, American robin, bulbul, jay, magpie, chickadee, American dipper, kite (bird of prey), bald eagle, vulture, great grey owl, fire salamander