# BEYOND INTERFERENCE CONTROL: RETAINING AND STRUCTURING KNOWLEDGE FOR PARAMETER-EFFICIENT CONTINUAL LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Parameter-Efficient Continual Learning (PECL) enables pretrained models to continually acquire new knowledge through lightweight adaptation, yet existing methods primarily focus on mitigating cross-task interference rather than explicitly controlling how knowledge is retained and organized throughout continual adaptation. In particular, the retention of shared knowledge is often governed by local update rules whose cumulative effect depends on the task optimization horizon, while architectural separation between shared and task-specific adaptations does not necessarily establish differentiated structural roles. To address these limitations, we propose Retaining and Structuring Knowledge (RSK), a general PECL framework that complements interference control through explicit knowledge retention and structural organization. RSK introduces Retention-Controlled Knowledge Consolidation with EMA (RKC-EMA), which parameterizes consolidation by a task-level cumulative retention target and derives the corresponding local EMA coefficient according to the optimization horizon. We theoretically establish that this formulation provides horizon-invariant cumulative retention and further characterize its effect on consolidated shared-state drift. RSK further introduces Structured Shared–Specific Knowledge Decomposition (SKD), which imposes asymmetric structural priors to encourage compact and historically stable shared adaptation while representing task-dependent variations through selective task-specific innovations. Extensive experiments on four continual-learning benchmarks demonstrate that RSK improves parameter-efficient continual adaptation and complements multiple representative PECL methods with marginal computational overhead.

## 1 INTRODUCTION

Continual Learning (CL) aims to enable models to continuously acquire new knowledge from evolving task or class streams while preserving previously learned capabilities. A central challenge is *catastrophic forgetting*: updates induced by new tasks can interfere with previously acquired knowledge (Saha et al., 2021; Qiu et al., 2025). Existing approaches mitigate such interference through regularization (Kirkpatrick et al., 2017), architectural isolation (Mallya & Lazebnik, 2018), experience replay (Mai et al., 2021), and gradient projection (Saha et al., 2021; Lin et al., 2022; Zhai et al., 2026). With the widespread adoption of pretrained models, however, continually updating the entire model becomes computationally and storage intensive and may compromise the reuse of pretrained knowledge (Hu et al., 2021; Yang et al., 2026). Parameter-Efficient Continual Learning (PECL) addresses this challenge by freezing most pretrained parameters and adapting only a small parameter subset (Wang et al., 2022b; Liang & Li, 2024; Qiao et al., 2024). In particular, low-rank adaptation (LoRA) provides a simple and effective parameterization for continual adaptation, making it a widely adopted basis for PECL.

Most existing PECL methods are primarily designed from the perspective of interference control. O-LoRA (Wang et al., 2023) constrains task-specific LoRA subspaces through orthogonality, while SD-LoRA (Wu et al., 2025) decouples adaptation directions and magnitudes. InfLoRA (Liang & Li, 2024) and SplitLoRA (Qiu et al., 2025) exploit historical representations or gradient subspaces to constrain subsequent updates, and EWC-LoRA (Zheng et al., 2026b) protects important parame-

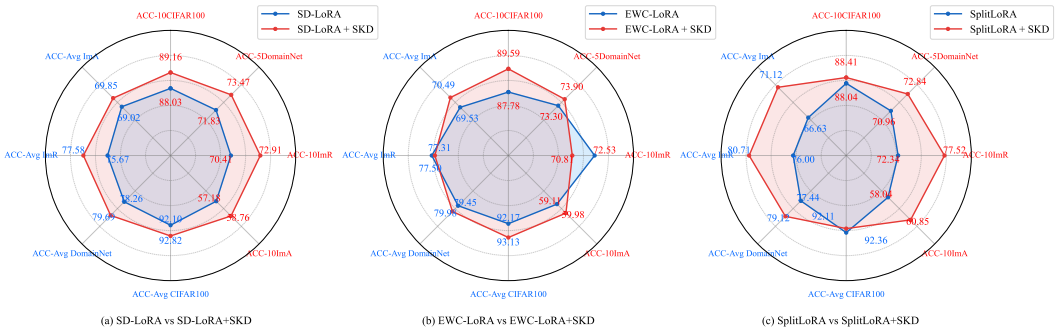


Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping 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.

ters through Fisher-based regularization. These approaches have substantially advanced parameter-efficient continual adaptation by restricting destructive updates. Nevertheless, controlling interference does not fully specify how knowledge should evolve throughout continual learning. In particular, two complementary aspects remain insufficiently controlled: *how previously consolidated shared knowledge should persist over an optimization trajectory*, and *how shared and task-specific adaptations should be structurally organized*.

The first issue arises from the distinction between local smoothing and trajectory-level retention. Exponential moving average (EMA) provides a natural mechanism for consolidating continually updated shared parameters and has been adopted in PECL-related architectures (Gao et al., 2024; Lou et al., 2026). However, a fixed local EMA coefficient $\alpha$ does not correspond to a fixed amount of retained knowledge over a complete task trajectory. After $K_t$ optimization steps, the contribution of the pre-task consolidated state is governed by $\alpha^{K_t}$; consequently, the same local coefficient induces different cumulative retention when the optimization horizon changes. This coupling makes the trajectory-level retention of shared knowledge implicit and horizon dependent. Rather than selecting a local smoothing coefficient and accepting its induced cumulative effect, a more direct formulation is to specify how much of the consolidated shared state should remain after learning a task and derive the local consolidation rule accordingly.

A second issue concerns the organization of shared and task-specific adaptation. Separating these components architecturally provides dedicated parameter pathways, but architectural separation alone does not determine their structural roles. Without additional structural guidance, the shared and task-specific branches may encode overlapping or redundant adaptations, weakening the intended distinction between reusable cross-task knowledge and task-dependent variations. At the same time, enforcing strict geometric separation can be unnecessarily restrictive because transferable and task-dependent adaptations need not be completely orthogonal. This motivates a softer structural principle: the shared pathway should preferentially capture compact and historically stable adaptation, while the task-specific pathway should retain sufficient flexibility to express selective task-dependent innovations.




(a) SD-LoRA vs SD-LoRA+SKD (b) EWC-LoRA vs EWC-LoRA+SKD (c) SplitLoRA vs SplitLoRA+SKD

Figure 1: Performance comparison of SD-LoRA (Wu et al., 2025), EWC-LoRA (Zheng et al., 2026b), and SplitLoRA (Qiu et al., 2025) with their RSK-enhanced variants across CIFAR-100, DomainNet, ImageNet-R, and ImageNet-A. The results provide an empirical overview of how RSK complements different PECL mechanisms across heterogeneous continual-learning benchmarks. ACC and Avg denote final and average accuracy, respectively (higher is better). Different scales are used across subfigures for visualization.

Motivated by these observations, we propose Retaining and Structuring Knowledge (RSK), a general framework that complements interference-oriented PECL through explicit knowledge retention and structural organization. RSK continuously consolidates a shared adaptation pathway while allowing its task-specific pathway to incorporate existing PECL update mechanisms. For knowledge retention, we introduce Retention-Controlled Knowledge Consolidation with EMA (RKC-EMA). Instead of directly fixing the local EMA coefficient, RKC-EMA parameterizes consolidation through a task-level cumulative retention target $R$. Given the task optimization horizon $K_t$, the corresponding local coefficient is determined as $\alpha_t = R^{1/K_t}$, such that the contribution of the pre-task consolidated state after the complete optimization trajectory remains $R$. We theoretically establish the resulting horizon-invariant cumulative retention and further characterize its effect on consolidated shared-state drift. For structural organization, we introduce Structured Shared–Specific Knowl-

edge Decomposition (SKD), which imposes asymmetric structural priors to encourage compact and historically stable shared adaptation while representing task-dependent variations through selective task-specific innovations. The two components operate complementarily: SKD structures online adaptation within each task, whereas RKC-EMA controls how the resulting shared knowledge is consolidated across task boundaries. Our main contributions are summarized as follows:

- We revisit parameter-efficient continual learning beyond the prevailing perspective of interference control and identify trajectory-level knowledge retention and shared–specific structural organization as two complementary aspects of continual adaptation. Based on this perspective, we introduce RSK, a general framework for retaining and structuring knowledge during PECL.
- We introduce RKC-EMA, which parameterizes consolidation through an explicit cumulative retention target and derives the corresponding local EMA coefficient according to the task optimization horizon. We establish horizon-invariant cumulative retention and characterize the resulting consolidated shared-state drift, providing a principled connection between trajectory-level retention and the stability–plasticity trade-off.
- We introduce SKD, which imposes asymmetric structural priors on shared and task-specific adaptations, encouraging compact and historically stable shared adaptation together with selective task-specific innovations. Extensive experiments across four continual-learning benchmarks further demonstrate the effectiveness of RSK and its compatibility with multiple representative PECL methods, with marginal computational overhead.

## 2 Related Work

### 2.1 Continual Learning

Continual Learning (CL) aims to enable models to continuously acquire and adapt knowledge from evolving task or class streams while preserving previously learned capabilities (Thrun, 1998; Zheng et al., 2026a). A fundamental challenge is catastrophic forgetting, where optimization on new tasks interferes with knowledge acquired from previous tasks. Existing CL methods have developed several major strategies to balance knowledge preservation and adaptation. Replay-based methods retain representative historical samples or their generated surrogates and revisit them during subsequent learning (Chaudhry et al., 2019; Buzzega et al., 2020). Architecture-based methods allocate or activate task-dependent parameters and modules, reducing interference by separating the parameter spaces used across tasks (Pietron et al., 2025). Regularization-based methods constrain parameter updates according to the importance of previously acquired knowledge (Li & Hoiem, 2016), while gradient-projection methods restrict new-task updates to directions that interfere less with historical task subspaces (Saha et al., 2021; Zhai et al., 2026). These paradigms provide different mechanisms for balancing stability on previously learned tasks with plasticity toward new knowledge. With the increasing adoption of large pretrained models, repeatedly updating and storing the entire model for sequential tasks becomes increasingly expensive. Parameter-Efficient Continual Learning (PECL) therefore combines continual learning with parameter-efficient adaptation by freezing most pretrained parameters and optimizing only a small parameter subset (Wang et al., 2022b; Liang & Li, 2024; Qiao et al., 2024). Among different parameter-efficient formulations, LoRA (Hu et al., 2021) represents model adaptation through low-rank parameter updates and has become a widely used basis for PECL due to its low trainable-parameter and storage overhead.

### 2.2 Parameter-Efficient Continual Learning

Recent PECL methods explore different parameter-efficient mechanisms for preserving previously acquired knowledge while adapting to sequential tasks. O-LoRA (Wang et al., 2023) encourages task-specific updates to occupy orthogonal subspaces, while InfLoRA (Liang & Li, 2024) constructs update directions based on historical representation subspaces. SD-LoRA (Wu et al., 2025) decouples low-rank adaptation into direction and magnitude components, and SplitLoRA (Qiu et al., 2025) organizes the parameter space according to stability and plasticity. EWC-LoRA (Zheng et al., 2026b) estimates parameter importance through Fisher information and regularizes subsequent updates accordingly. Beyond task-specific update constraints, recent studies have also explored shared or temporally consolidated adaptation. EMA-LoRA (Gao et al., 2024) employs exponential

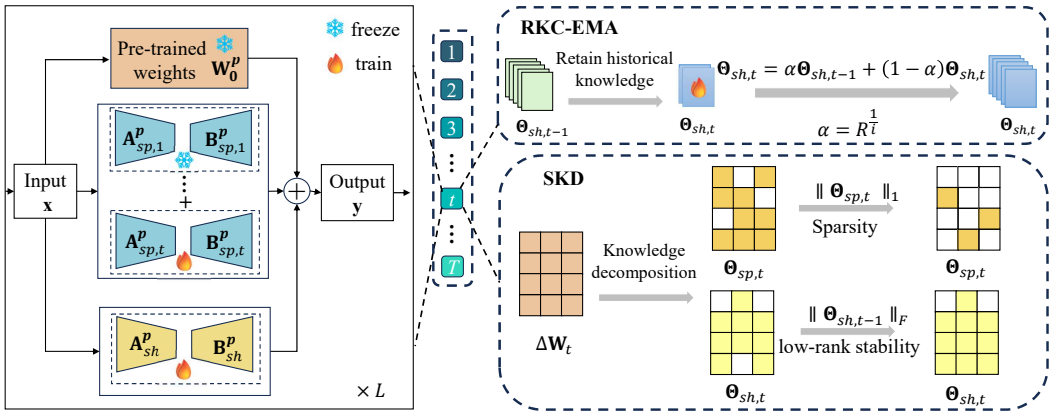



Figure 2: **Overview of RSK.** A frozen pre-trained model is augmented with shared and task-specific LoRA pathways. SKD structures online adaptation through asymmetric priors, while RKC-EMA consolidates the shared trajectory under a task-level retention target and propagates it across tasks.

moving averages to stabilize LoRA updates, while Lou et al. (2026) combines an EMA-updated shared adapter with task-specific experts and bi-level routing. Together, these studies span subspace constraints, parameter regularization, shared adaptation, and temporal consolidation, reflecting the growing range of parameter-efficient mechanisms for continual knowledge preservation and adaptation.

# 3 Method

## 3.1 Problem Formulation and Overview

We consider PECL over a sequence of tasks $\{\mathcal{T}_t\}_{t=1}^{T}$, where $t$ indexes the current task and $T$ denotes the total number of tasks. Following LoRA-based PEFT (Hu et al., 2021; Qiu et al., 2025), we adapt the key and value projections, indexed by $p \in \{k, v\}$, while keeping the pre-trained backbone frozen. For projection $p$, let $\mathbf{W}_0^p \in \mathbb{R}^{d\times d}$ denote the frozen pre-trained weight and $\mathbf{W}_t^p$ its adapted weight for task $\mathcal{T}_t$. We decompose the adaptation into a shared update $\Delta\mathbf{W}^p_{\mathrm{sh},t}$ and a task-specific update $\Delta\mathbf{W}^p_{\mathrm{sp},t}$:

$$\begin{aligned}\mathbf{W}_t^p &= \mathbf{W}_0^p + \Delta\mathbf{W}^p_{\mathrm{sh},t} + \Delta\mathbf{W}^p_{\mathrm{sp},t} \\ &= \mathbf{W}_0^p + \mathbf{B}^p_{\mathrm{sh},t}\mathbf{A}^p_{\mathrm{sh},t} + \mathbf{B}^p_{\mathrm{sp},t}\mathbf{A}^p_{\mathrm{sp},t},\end{aligned} \tag{1}$$

where $\mathbf{A}^p_{\mathrm{sh},t}, \mathbf{A}^p_{\mathrm{sp},t} \in \mathbb{R}^{r\times d}$ and $\mathbf{B}^p_{\mathrm{sh},t}, \mathbf{B}^p_{\mathrm{sp},t} \in \mathbb{R}^{d\times r}$, with LoRA rank $r \ll d$. A complete notation summary is provided in Appendix A.1.

The central problem is how to continually update $\Delta\mathbf{W}^p_{\mathrm{sh},t}$ and $\Delta\mathbf{W}^p_{\mathrm{sp},t}$ while controlling both *temporal retention* and *structural allocation*. Let $K_t$ denote the number of optimization steps for task $\mathcal{T}_t$. First, the shared update $\Delta\mathbf{W}^p_{\mathrm{sh},t}$ should retain a controlled contribution from previously consolidated knowledge even when $K_t$ varies across tasks. Second, the decomposition in Eq. 1 alone does not determine how newly learned adaptation should be allocated between $\Delta\mathbf{W}^p_{\mathrm{sh},t}$ and $\Delta\mathbf{W}^p_{\mathrm{sp},t}$, leaving their structural roles insufficiently differentiated.

To address these two challenges, we propose Retaining and Structuring Knowledge (RSK), illustrated in Fig. 2. For each task, SKD first imposes asymmetric structural priors on the shared and task-specific updates, encouraging compact and historically consistent shared adaptation together with selective task-specific variation. RKC-EMA then consolidates the resulting shared optimization trajectory by deriving the local EMA coefficient from a prescribed task-level retention target and propagates the consolidated shared state to the next task.

### 3.2 RETENTION-CONTROLLED KNOWLEDGE CONSOLIDATION

**Motivation.** The shared adaptation is continually updated with current-task information using EMA (Gao et al., 2024; Lou et al., 2026), where the local decay coefficient $\alpha_t$ determines the retention of previously consolidated knowledge at each optimization step of task $\mathcal{T}_t$. However, the cumulative retention over a task is jointly determined by $\alpha_t$ and the optimization horizon $K_t$. Consequently, the same local coefficient can yield different retention levels across tasks with different optimization horizons. To achieve consistent retention across tasks, we instead specify a task-level retention target and derive the corresponding local coefficient.

**Key Design.** Let $\mathbf{\Theta}_{\mathrm{sh},t,i}$ denote the collection of online shared LoRA factors after optimization step $i$ of task $\mathcal{T}_t$, and let $\bar{\mathbf{\Theta}}_{\mathrm{sh},t,i}$ denote the corresponding consolidated state, where $i \in \{0, \ldots, K_t\}$ indexes within-task optimization steps. For $t \geq 2$, both states are initialized from the shared state consolidated after task $\mathcal{T}_{t-1}$. Over the $K_t$ optimization steps, the consolidation trajectory is

$$\begin{aligned} \bar{\mathbf{\Theta}}_{\mathrm{sh},t,i} &= \alpha_t \bar{\mathbf{\Theta}}_{\mathrm{sh},t,i-1} + (1-\alpha_t)\mathbf{\Theta}_{\mathrm{sh},t,i}, \\ \bar{\mathbf{\Theta}}_{\mathrm{sh},t,K_t} &= \alpha_t^{K_t} \bar{\mathbf{\Theta}}_{\mathrm{sh},t,0} + (1-\alpha_t) \sum_{i=1}^{K_t} \alpha_t^{K_t-i} \mathbf{\Theta}_{\mathrm{sh},t,i}, \qquad R_t \triangleq \alpha_t^{K_t}. \end{aligned} \tag{2}$$

All operations are applied component-wise to the shared LoRA factors. Eq. 2 makes explicit that $\alpha_t$ is a *local* consolidation coefficient, whereas $R_t = \alpha_t^{K_t}$ is the exact coefficient of the pre-task consolidated state after the complete optimization trajectory. RKC-EMA therefore reverses the conventional parameterization: rather than directly selecting $\alpha_t$, we specify a target cumulative retention $R \in (0,1)$ and derive the local coefficient required to realize $R_t = R$. Here, *cumulative retention* refers specifically to the contribution of the pre-task consolidated parameter state to the post-task consolidated state. It is therefore an explicit property of the consolidation dynamics, rather than a direct measure of semantic or task-level knowledge retention.

**Theorem 1 (Horizon-Invariant Cumulative Retention).** For any target cumulative retention $R \in (0,1)$ and optimization horizon $K_t \geq 1$, there exists a unique constant EMA coefficient

$$\alpha_t = R^{1/K_t}, \tag{3}$$

such that the realized cumulative retention satisfies $R_t = \alpha_t^{K_t} = R$. Consequently, for arbitrary tasks $\mathcal{T}_a$ and $\mathcal{T}_b$, the retention and drift properties satisfy

$$\begin{aligned} \alpha_a^{K_a} = \alpha_b^{K_b} = R, \qquad K_a \neq K_b, \\ R_{\mathrm{fixed},t} = \alpha^{K_t}, \qquad \alpha \in (0,1), \\ \left\|\bar{\boldsymbol{\theta}}_{\mathrm{sh},t,K_t} - \bar{\boldsymbol{\theta}}_{\mathrm{sh},t,0}\right\|_F \leq (1-R)\delta_t, \qquad \text{if} \quad \max_{1\leq i\leq K_t} \left\|\boldsymbol{\theta}_{\mathrm{sh},t,i} - \bar{\boldsymbol{\theta}}_{\mathrm{sh},t,0}\right\|_F \leq \delta_t. \end{aligned} \tag{4}$$

Here, $\boldsymbol{\theta}_{\mathrm{sh},t,i} \in \mathbf{\Theta}_{\mathrm{sh},t,i}$ denotes an individual shared LoRA factor, and $\bar{\boldsymbol{\theta}}_{\mathrm{sh},t,i}$ denotes its consolidated counterpart. The bound is applied component-wise to each shared factor.

*Proof sketch.* Unrolling the EMA recurrence in Eq. 2 gives the coefficient $\alpha_t^{K_t}$ for the pre-task consolidated state. Setting this coefficient to $R$ yields $\alpha_t = R^{1/K_t}$; uniqueness follows from the strict monotonicity of $\alpha^{K_t}$ on $(0,1)$. The drift inequality follows by subtracting $\bar{\boldsymbol{\theta}}_{\mathrm{sh},t,0}$ from the unrolled recurrence and applying the triangle inequality. The complete derivation is provided in Appendix A.2.1.

Theorem 1 establishes the distinction between local smoothing and trajectory-level retention. Under a fixed local coefficient, the contribution of the pre-task state varies exponentially with the optimization horizon. Under RKC-EMA, $R$ retains the same parameter-level interpretation across tasks even when their horizons differ. The drift bound further characterizes how this retention target constrains consolidated parameter evolution: larger $R$ yields a tighter upper bound on deviation from the pre-task state, whereas smaller $R$ permits greater adaptation. RKC-EMA therefore exposes the stability–plasticity trade-off directly at the task-trajectory level.

### 3.3 Structured Shared–Specific Knowledge Decomposition

**Motivation.** RKC-EMA controls how historical shared knowledge is retained, but does not determine how current adaptation should be allocated between the shared and task-specific pathways. Existing PECL methods commonly reduce interference through architectural separation or constrained adaptation (Wu et al., 2025; Qiu et al., 2025). However, architectural separation alone may still yield redundant updates, while strict geometric separation can be overly restrictive when transferable and task-dependent variations contain correlated directions. We therefore differentiate the two pathways through asymmetric structural priors rather than hard separation.

**Key Design.** The shared pathway is encouraged to remain compact and close to its consolidated state, while the task-specific pathway captures selective innovations through sparsity on $\mathbf{B}^p_{\mathrm{sp},t}$, leaving $\mathbf{A}^p_{\mathrm{sp},t}$ as a flexible adaptation basis. Let $\mathcal{S}_{\mathrm{sh},t}$ denote the collection of shared LoRA factors and $\mathcal{S}^B_{\mathrm{sp},t}$ the collection of task-specific output factors. For each $\boldsymbol{\theta} \in \mathcal{S}_{\mathrm{sh},t}$, let $\bar{\boldsymbol{\theta}}$ denote its consolidated counterpart. SKD is defined as

$$\begin{aligned}\mathcal{L}_{\mathrm{SKD},t} = \frac{\lambda_{\mathrm{cmp}}}{|\mathcal{S}_{\mathrm{sh},t}|} \sum_{\boldsymbol{\theta}\in\mathcal{S}_{\mathrm{sh},t}} \|\boldsymbol{\theta}\|_F + \frac{\lambda_{\mathrm{fid}}}{|\mathcal{S}_{\mathrm{sh},t}|} \sum_{\boldsymbol{\theta}\in\mathcal{S}_{\mathrm{sh},t}} \|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_F^2 \\ + \frac{\lambda_{\mathrm{sp}}}{|\mathcal{S}^B_{\mathrm{sp},t}|} \sum_{\mathbf{B}\in\mathcal{S}^B_{\mathrm{sp},t}} \|\mathbf{B}\|_1,\end{aligned} \tag{5}$$

where $\lambda_{\mathrm{cmp}}$, $\lambda_{\mathrm{fid}}$, and $\lambda_{\mathrm{sp}}$ control shared compactness, historical fidelity, and task-specific sparsity, respectively.

**Structural Interpretation.** The three terms in Eq. 5 impose complementary structural preferences aligned with the intended roles of the two pathways. For shared compactness, the factor regularization controls an upper bound on the nuclear complexity of the induced update, since $\|\Delta\mathbf{W}^p_{\mathrm{sh},t}\|_* \le \|\mathbf{B}^p_{\mathrm{sh},t}\|_F \|\mathbf{A}^p_{\mathrm{sh},t}\|_F$, thereby discouraging excessive task-local variation in the shared pathway. For historical fidelity, the negative gradient of $\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_F^2$ points toward the consolidated state $\bar{\boldsymbol{\theta}}$, anchoring online shared adaptation to previously accumulated knowledge. For task-specific sparsity, the $\ell_1$ prior admits $[\mathbf{B}^p_{\mathrm{sp},t}]_{ij} = 0$ as a stationary solution when the corresponding task-gradient magnitude is sufficiently small, promoting selective rather than redundant task-specific variation. Together, SKD encourages *compact and historically consistent shared adaptation* and *selective task-specific innovation*, without requiring strict orthogonality or semantic disentanglement. Complete derivations are provided in Appendix A.2.2.

### 3.4 Unified Optimization

RSK follows a task-wise inherit–adapt–consolidate–propagate procedure. For each incoming task, the shared pathway inherits the knowledge consolidated from previous tasks, while the current task-specific pathway is initialized for new adaptation. SKD then structures the online update by encouraging compact and historically consistent shared adaptation together with selective task-specific variation. After each optimization step, RKC-EMA consolidates the updated shared state using the task-calibrated coefficient $\alpha_t = R^{1/K_t}$. The resulting consolidated state is finally propagated to the next task, thereby coupling within-task adaptation with cross-task knowledge retention.

**Procedure.** Let $\boldsymbol{\Theta}_{\mathrm{sp},t,i}$ denote the task-specific LoRA state after step $i$ of task $\mathcal{T}_t$, and let $\mathrm{OptStep}(\cdot)$ denote one update of the adopted optimizer. For each task $t \ge 2$, RSK performs

$$\begin{aligned}
&\textit{Initialization:} && \boldsymbol{\Theta}_{\mathrm{sh},t,0} \leftarrow \bar{\boldsymbol{\Theta}}_{\mathrm{sh},t-1,K_{t-1}}, \quad \bar{\boldsymbol{\Theta}}_{\mathrm{sh},t,0} \leftarrow \bar{\boldsymbol{\Theta}}_{\mathrm{sh},t-1,K_{t-1}}, \quad \alpha_t \leftarrow R^{1/K_t}, \\
&\textit{Online adaptation:} && \mathcal{L}_{t,i} = \mathcal{L}_{\mathrm{CL},t,i} + \mathcal{L}_{\mathrm{SKD},t,i}, \\
&&& (\boldsymbol{\Theta}_{\mathrm{sh},t,i}, \boldsymbol{\Theta}_{\mathrm{sp},t,i}) \leftarrow \mathrm{OptStep}\left(\boldsymbol{\Theta}_{\mathrm{sh},t,i-1}, \boldsymbol{\Theta}_{\mathrm{sp},t,i-1}; \mathcal{L}_{t,i}\right), \quad i = 1, \dots, K_t, \\
&\textit{Consolidation:} && \bar{\boldsymbol{\Theta}}_{\mathrm{sh},t,i} \leftarrow \alpha_t \bar{\boldsymbol{\Theta}}_{\mathrm{sh},t,i-1} + (1-\alpha_t)\boldsymbol{\Theta}_{\mathrm{sh},t,i}, \\
&\textit{Propagation:} && \boldsymbol{\Theta}_{\mathrm{sh},t+1,0} \leftarrow \bar{\boldsymbol{\Theta}}_{\mathrm{sh},t,K_t}.
\end{aligned} \tag{6}$$

For the first task $\mathcal{T}_1$, no historical consolidated state is available; its learned shared state therefore initializes subsequent continual consolidation. For $t \geq 2$, Eq. 6 recursively connects historical-state inheritance, SKD-regularized online adaptation, RKC-EMA consolidation, and cross-task propagation. The complete implementation is provided in Algorithm 3 of Appendix A.3.

# 4 EXPERIMENT

## 4.1 EXPERIMENTAL SETTING

**Datasets.** We conduct experiments on **four** standard CL benchmarks: CIFAR-100 (Krizhevsky et al., 2009), which contains 100 classes divided into 10 tasks with 10 classes per task; DomainNet (Peng et al., 2019), which contains 345 classes divided into 5 tasks with 69 classes per task; ImageNet-R (Hendrycks et al., 2021), which comprises 200 classes of stylized rendered images divided into 10 tasks with 20 classes per task; and ImageNet-A (Hendrycks et al., 2021), which comprises 200 classes of natural adversarial examples divided into 10 tasks with 20 classes per task.

**Baselines and Evaluation Metrics.** We compare our method with several state-of-the-art CL approaches, including Vanilla LoRA (Hu et al., 2021), L2P (Wang et al., 2022b), DualPrompt (Wang et al., 2022a), CODA-Prompt (Smith et al., 2023), O-LoRA (Wang et al., 2023), InfLoRA (Liang & Li, 2024), SD-LoRA (Wu et al., 2025), CL-LoRA (He et al., 2025), BiLoRA (Zhu et al., 2025), EWC-LoRA (Zheng et al., 2026b), SplitLoRA (Qiu et al., 2025). We report final accuracy, average accuracy, forgetting, and stability, whose definitions are provided in Appendix A.4.

**Implementation Details.** All experiments are conducted on a single NVIDIA A40 46G GPU. Following Qiu et al. (2025), Vanilla LoRA, and RSK use batch size 256, 10 epochs, lr $1 \times 10^{-3}$, classifier lr $1 \times 10^{-2}$, and AdamW. Other methods follow their original configurations (Zheng et al., 2026b). The cumulative retention target of RKC-EMA is set to $R = 0.5$. The regularization strength follows the setting adopted in prior work (Wang et al., 2023; Liu et al., 2026), with candidate values ranging from $10^{-7}$ to $10^{-1}$. Results are averaged over three random seeds (1993, 1997, and 2026).

Table 1: Comparison results on CIFAR-100, DomainNet, ImageNet-R, and ImageNet-A (in %). **Bold** and underline indicate the highest and second-highest scores, respectively.

| Tasks | CIFAR-100 | | DomainNet | | ImageNet-R | | ImageNet-A | |
|---|---|---|---|---|---|---|---|---|
| Methods | $\bar{A}_{10}$ (↑) | Avg. (↑) | $\bar{A}_5$ (↑) | Avg. (↑) | $\bar{A}_{10}$ (↑) | Avg. (↑) | $\bar{A}_{10}$ (↑) | Avg. (↑) |
| Joint Train | 92.82±0.20 | 95.41±0.05 | 76.84±0.06 | 81.25±0.05 | 81.69±0.28 | 86.25±0.09 | 65.01±0.74 | 74.10±0.44 |
| Finetune | 79.09±1.53 | 88.17±0.45 | 65.57±0.20 | 75.12±0.12 | 60.42±1.64 | 73.18±0.32 | 32.85±1.53 | 54.55±1.44 |
| L2P | 83.18±1.20 | 87.69±1.05 | 70.26±0.25 | 75.83±0.98 | 71.26±0.44 | 76.13±0.46 | 42.94±1.27 | 51.40±1.95 |
| DualPrompt | 81.48±0.86 | 86.41±0.66 | 68.26±0.90 | 73.84±0.45 | 68.22±0.20 | 73.81±0.39 | 45.49±0.96 | 54.68±1.24 |
| O-LoRA | 87.03±0.55 | 91.45±0.24 | 66.51±1.43 | 75.69±0.56 | 75.96±0.23 | 79.66±0.28 | 58.18±0.55 | 68.46±0.55 |
| CODA-Prompt | 86.31±0.12 | 90.67±0.22 | 70.58±0.53 | 76.68±0.44 | 74.05±0.41 | 78.14±0.39 | 45.36±0.78 | 57.03±0.94 |
| CL-LoRA | 87.65±0.53 | 92.25±0.41 | 71.06±0.31 | 77.76±0.29 | **78.72±0.44** | **85.20±0.45** | 57.62±0.89 | 70.76±0.63 |
| InfLoRA | 85.98±0.49 | 91.40±0.12 | 72.65±0.02 | 79.08±0.06 | 75.38±0.24 | 80.03±0.08 | 50.00±0.74 | 64.31±0.44 |
| BiLoRA | 85.99±0.49 | 90.62±0.42 | 69.75±0.23 | 73.86±0.21 | 74.28±0.92 | 77.38±0.88 | 51.05±0.74 | 62.82±0.65 |
| SD-LoRA | 88.03±0.08 | 92.10±0.04 | 71.83±0.04 | 78.26±0.09 | 70.41±0.46 | 75.67±0.25 | 57.18±0.44 | 69.02±0.32 |
| SplitLoRA | 88.04±0.06 | 92.36±0.10 | 70.96±0.36 | 77.44±0.36 | 72.34±0.35 | 76.00±0.44 | 58.04±0.71 | 66.63±1.63 |
| EWC-LoRA | 87.78±0.31 | 92.17±0.10 | 73.30±0.14 | 79.45±0.09 | 72.53±0.38 | 77.50±0.18 | 59.11±0.56 | 69.53±0.17 |
| Vanilla LoRA | 85.45±0.67 | 90.53±0.44 | 68.72±0.16 | 76.52±0.09 | 75.52±0.42 | 79.43±0.36 | 57.78±0.90 | 68.78±0.42 |
| RSK (Ours) | 87.95±0.27 | 91.87±0.68 | 73.27±0.01 | 79.39±0.05 | 76.71±0.35 | 80.19±0.43 | 59.71±0.39 | 69.90±0.45 |
| SD-LoRA+RSK | 89.16±0.07 | 92.82±0.11 | 73.47±0.07 | 79.69±0.02 | 72.91±0.32 | 77.58±0.05 | 58.76±1.13 | 69.85±0.46 |
| SplitLoRA+RSK | 88.41±0.29 | 92.11±0.21 | 72.84±0.29 | 79.12±0.14 | 77.52±0.13 | 80.71±0.08 | **60.85±0.25** | **71.12±0.35** |
| EWC-LoRA+RSK | **89.59±0.16** | **93.13±0.12** | **73.90±0.08** | **79.98±0.06** | 70.81±0.26 | 77.31±0.33 | 59.98±0.21 | 70.49±0.46 |

## 4.2 MAIN RESULT

**Comparison with Various CL Baselines.** We compare RSK with a broad range of continual learning and parameter-efficient continual learning methods, as shown in Tab. 1. RSK consistently improves upon Vanilla LoRA across all four benchmarks, with gains of 2.50, 4.55, 1.19, and 1.93 percentage points on CIFAR-100, DomainNet, ImageNet-R, and ImageNet-A, respectively. Notably, RSK achieves 73.27% on DomainNet, outperforming InfLoRA (72.65%), and reaches 76.71% on ImageNet-R, surpassing InfLoRA (75.38%) and BiLoRA (74.28%), despite requiring neither historical task-specific parameters nor accumulated subspaces or Fisher statistics. These improvements are particularly evident on DomainNet and ImageNet-R, suggesting that explicitly retaining and structuring shared knowledge is beneficial under heterogeneous and distribution-varying continual-learning streams. Furthermore, RSK can be readily combined with existing PECL methods, yield-

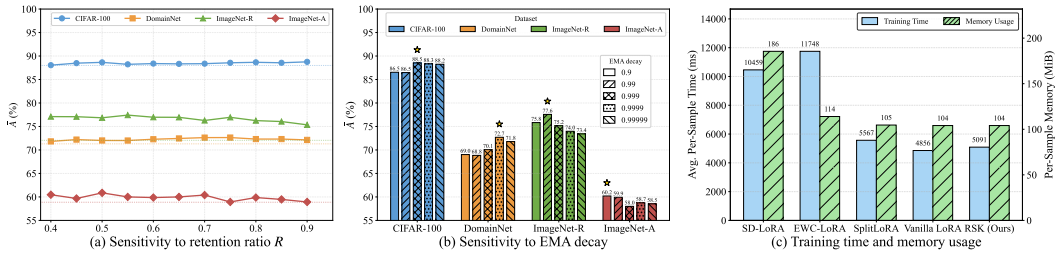



Figure 3: Sensitivity analysis of RKC-EMA. (a) Retention ratio $R$ (0.40–0.90); dotted lines denote the base method. (b) EMA decay factors (0.9–0.99999); stars indicate the best decay per dataset. (c) Per-sample training time and memory usage normalized by batch size.

ing consistent gains across different adaptation mechanisms. For example, EWC-LoRA+RSK improves over EWC-LoRA by 1.81 points on CIFAR-100 and 0.60 points on DomainNet, while SplitLoRA+RSK improves ImageNet-R by 5.18 points and ImageNet-A by 2.81 points. Similar improvements across different PECL backbones indicate that RSK does not depend on a particular interference-control strategy, but instead provides a complementary mechanism for retaining and organizing shared knowledge. Overall, these results support the central motivation of RSK: effective continual adaptation benefits not only from suppressing interference, but also from explicitly controlling how knowledge is retained and structured across tasks.

**Sensitivity Comparison of EMA and RKC-EMA.** As shown in Fig. 3(a), RKC-EMA is stable across a wide range of retention ratios $R$, and a single $R$ achieves near-optimal accuracy on all four datasets. For example, the accuracy varies only within 88.05–88.76 on CIFAR-100 and 58.93–60.87 on ImageNet-A, with most settings outperforming the base method. In contrast, as shown in Fig. 3(b), the optimal EMA decay varies substantially across datasets (0.999 on CIFAR-100, 0.9999 on DomainNet and ImageNet-A, and 0.99 on ImageNet-R), indicating that a fixed EMA decay must be tuned per dataset and does not transfer across benchmarks. Full results are reported in Appendix A.5. This highlights the advantage of task-level retention control over direct decay-factor selection.

**Memory Footprint and Training Time.** As shown in Fig. 3(c), RSK incurs memory and training time comparable to Vanilla LoRA, yet substantially less training time than SD-LoRA and EWC-LoRA. This indicates that RSK improves continual learning performance with negligible computational overhead.

**Analysis of Plasticity and Stability.** As shown in Fig.4, compared with Vanilla LoRA, RSD substantially improves stability and reduces forgetting while achieving higher total accuracy, demonstrating its effectiveness in balancing stability and plasticity. When combined with existing PECL methods, RSD further improves stability for SD-LoRA and SplitLoRA, while its effect on EWC-LoRA is relatively different, with improvements mainly reflected in overall accuracy.

Table 2: Final accuracy ($A_k$) on ImageNet-A after training on $k \in \{5, 10, 20\}$ tasks.

| Methods | $A_5$ | $A_{10}$ | $A_{20}$ |
|---|---|---|---|
| SD-LoRA | 59.73 | 56.80 | **53.60** |
| SD-LoRA + RSK | **60.67** | **59.20** | 52.33 |
| EWC-LoRA | 59.87 | 58.33 | 55.27 |
| EWC-LoRA + RSK | **62.47** | **60.27** | **56.87** |
| SplitLoRA | 59.47 | 58.87 | 47.27 |
| SplitLoRA + RSK | **61.53** | **61.20** | **52.20** |
| Vanilla LoRA | 57.00 | 59.00 | 50.20 |
| RSK (Ours) | **60.67** | **60.27** | **52.47** |

**Results Across Varied Task Lengths.** Tab. 2 reports the final accuracy on Image-A under 5-, 10-, and 20-task settings. RSK improves most base methods, e.g., raising EWC-LoRA, SplitLoRA, and Vanilla LoRA by 2.60, 2.06, and 3.67 points at 5 tasks, and the gains remain positive at 10 and 20 tasks. A notable exception is SD-LoRA at 20 tasks, where RSK slightly reduces the accuracy from 53.60 to 52.33, suggesting that on longer sequences RSK may interfere with SD-LoRA's adaptation.

### 4.3 Ablation Study

To evaluate the contribution of each component, we conduct ablation studies across four representative PECL methods, as shown in Tab. 3. RKC-EMA provides the dominant gains across most

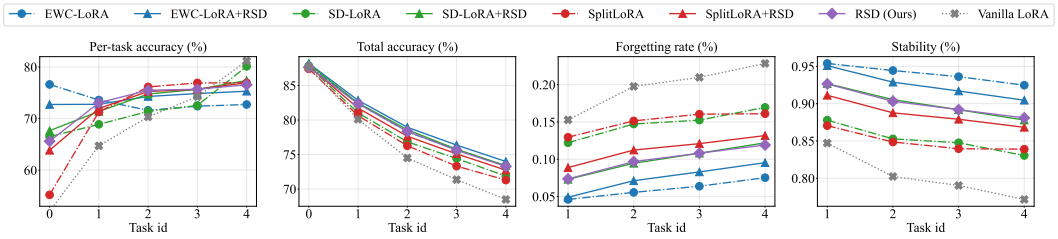



Figure 4: Comparison of per-task accuracy, total accuracy, forgetting rate, and stability on DomainNet across different methods. RSD achieves a favorable stability–plasticity trade-off when combined with existing state-of-the-art methods.

settings, confirming the importance of explicitly controlling shared-state retention. For example, it improves Vanilla LoRA on DomainNet from 68.72 to 72.22 (+3.50 points) and SplitLoRA on ImageNet-R from 72.34 to 76.99 (+4.65 points). With SKD further incorporated, the full RSK framework reaches gains of +4.55 and +5.18 points, respectively. SKD contributes smaller but complementary improvements by further regulating how new adaptation is distributed between the shared and task-specific pathways. For instance, it raises Vanilla LoRA on DomainNet from 72.22 to 73.27 (+1.05) and SplitLoRA on ImageNet-R from 76.99 to 77.52 (+0.53). Although minor fluctuations occur in a few settings, the overall trend is clear: RKC-EMA primarily improves *temporal retention*, while SKD provides complementary *structural allocation*. Together, they support the two core design objectives of RSK. This complementary behavior also explains why RSK remains effective across different PECL backbones rather than depending on a single adaptation mechanism.

Table 3: Ablation studies on CIFAR-100, DomainNet, ImageNet-R, and ImageNet-A (in %). "+ RSK" denotes combining the base method with both RKC-EMA and SKD. **Bold** and underline indicate the highest and second-highest scores, respectively.

| Tasks | CIFAR-100 | | DomainNet | | ImageNet-R | | ImageNet-A | |
|---|---|---|---|---|---|---|---|---|
| Methods | $\bar{A}_{10}$ (↑) | Avg. (↑) | $\bar{A}_5$ (↑) | Avg. (↑) | $\bar{A}_{10}$ (↑) | Avg. (↑) | $\bar{A}_{10}$ (↑) | Avg. (↑) |
| SD-LoRA | 88.03±0.08 | 92.10±0.04 | 71.83±0.04 | 78.26±0.09 | 70.41±0.46 | 75.67±0.25 | 57.18±0.44 | 69.02±0.32 |
| + RKC-EMA | 89.07±0.05 | **92.83±0.12** | 73.43±0.15 | 79.64±0.05 | 72.47±0.35 | 77.56±0.05 | 58.40±1.37 | 69.73±0.44 |
| + RSK | **89.16±0.07** | 92.82±0.11 | **73.47±0.07** | **79.69±0.02** | **72.91±0.32** | **77.58±0.05** | **58.76±1.13** | **69.85±0.46** |
| SplitLoRA | 88.04±0.06 | 92.36±0.10 | 70.96±0.36 | 77.44±0.36 | 72.34±0.35 | 76.00±0.44 | 58.04±0.71 | 66.63±1.63 |
| + RKC-EMA | 88.37±0.34 | 92.10±0.13 | 72.14±0.11 | 78.76±0.05 | 76.99±0.11 | 80.49±0.22 | 60.85±0.14 | 70.85±0.57 |
| + RSK | **88.41±0.29** | **92.11±0.21** | **72.84±0.29** | **79.12±0.14** | **77.52±0.13** | **80.71±0.08** | **60.85±0.25** | **71.12±0.35** |
| EWC-LoRA | 87.78±0.31 | 92.17±0.10 | 73.30±0.14 | 79.45±0.09 | **72.53±0.38** | **77.50±0.18** | 59.11±0.56 | 69.53±0.17 |
| + RKC-EMA | **89.63±0.13** | 93.09±0.08 | 73.46±0.12 | 79.82±0.05 | 70.76±0.01 | 77.25±0.35 | 59.60±0.09 | 70.38±0.45 |
| + RSK | 89.59±0.16 | **93.13±0.12** | **73.90±0.08** | **79.98±0.06** | 70.81±0.26 | 77.31±0.33 | **59.98±0.21** | **70.49±0.46** |
| Vanilla LoRA | 85.45±0.67 | 90.53±0.44 | 68.72±0.16 | 76.52±0.09 | 75.52±0.42 | 79.43±0.36 | 57.78±0.90 | 68.78±0.42 |
| + RKC-EMA | 87.87±0.26 | 91.80±0.65 | 72.22±0.09 | 78.80±0.02 | 76.57±0.36 | 79.95±0.46 | 59.20±0.33 | 69.88±0.45 |
| + RSK | **87.95±0.27** | **91.87±0.68** | **73.27±0.01** | **79.39±0.05** | **76.71±0.35** | **80.19±0.43** | **59.71±0.39** | **69.90±0.45** |

## 5 Conclusion

In this paper, we propose RSK, a knowledge organization framework for PECL. Unlike existing methods that primarily rely on historical task information to mitigate interference, RSK focuses on knowledge retention and organization, continually consolidating cross-task shared knowledge and promoting a structured and complementary division of roles between shared and task-specific knowledge without explicitly maintaining historical task-specific parameters. Specifically, we propose Retention-Controlled Knowledge Consolidation with EMA (RKC-EMA), which dynamically determines the EMA decay factor according to the task optimization horizon under a predefined cumulative retention target, thereby enabling stable control over the continual consolidation of shared knowledge. We further introduce the Structured Shared-Specific Knowledge Decomposition (SKD) loss, which encourages complementary roles between the shared and task-specific branches through asymmetric structural priors. Extensive experiments on four benchmark datasets and across multiple PECL methods demonstrate the effectiveness and compatibility of RSK, while further ablation and sensitivity analyses validate the effectiveness of both RKC-EMA and SKD.

### AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing of the manuscript. We have not used generative AI tools for retrieval and discovery, research ideation or execution, drafting sections of the paper, generating synthetic datasets, proving mathematical claims, or any other purpose. All AI-assisted writing was reviewed, revised, and verified by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

## REFERENCES

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.

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

Zijian Gao, Xingxing Zhang, Kele Xu, Xinjun Mao, and Huaimin Wang. Stabilizing zero-shot prediction: A novel antidote to forgetting in continual vision-language tasks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=C4zmR2kyP8.

Jiangpeng He, Zhihao Duan, and Fengqing Zhu. CL-LoRA: Continual low-rank adaptation for rehearsal-free class-incremental learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30534–30544, 2025.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8349, 2021.

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, 2021. URL https://arxiv.org/abs/2106.09685.

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

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 614–629, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46493-0.

Yan-Shuo Liang and Wu-Jun Li. Inflora: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23638–23647, 2024.

Sen Lin, Li Yang, Deliang Fan, and Junshan Zhang. Trgp: Trust region gradient projection for continual learning, 2022. URL https://arxiv.org/abs/2202.02931.

Shangge Liu, Yuehan Yin, Lei Wang, Qi Fan, Yinghuan Shi, Wenbin Li, Yang Gao, and Dacheng Tao. Understanding and enforcing weight disentanglement in task arithmetic, 2026. URL https://arxiv.org/abs/2604.17078.

Meng Lou, Yunxiang Fu, and Yizhou Yu. Scaling continual learning to 300+ tasks with bi-level routing mixture-of-experts, 2026. URL https://arxiv.org/abs/2602.03473.

Zheda Mai, Ruiwen Li, Hyunwoo Kim, and Scott Sanner. Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3589–3599, June 2021.

Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7765–7773, 2018. doi: 10.1109/CVPR.2018.00810.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1406–1415, 2019.

Marcin Pietron, Kamil Faber, Dominik Żurek, and Roberto Corizzo. Tinysubnets: An efficient and low capacity continual learning strategy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(19):19913–19920, Apr. 2025. doi: 10.1609/aaai.v39i19.34193. URL https://ojs.aaai.org/index.php/AAAI/article/view/34193.

Jingyang Qiao, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yong Peng, and Yuan Xie. Prompt gradient projection for continual learning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 16294–16315, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/46c2a9a6f2b2be68682013eb1173c801-Paper-Conference.pdf.

Haomiao Qiu, Miao Zhang, Ziyue Qiao, Weili Guan, Min Zhang, and Liqiang Nie. Splitlora: Balancing stability and plasticity in continual learning through gradient space splitting. *CoRR*, abs/2505.22370, May 2025. URL https://doi.org/10.48550/arXiv.2505.22370.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning, 2021. URL https://arxiv.org/abs/2103.09762.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11909–11919, 2023.

Sebastian Thrun. *Lifelong learning algorithms*, pp. 181–209. Kluwer Academic Publishers, USA, 1998. ISBN 0792380479.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuan-Jing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10658–10671, 2023.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European conference on computer vision*, pp. 631–648. Springer, 2022a.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 139–149, 2022b.

Yichen Wu, Hongming Piao, Long-Kai Huang, Renzhen Wang, Wanhua Li, Hanspeter Pfister, Deyu Meng, Kede Ma, and Ying Wei. Sd-lora: Scalable decoupled low-rank adaptation for class incremental learning. *arXiv preprint arXiv:2501.13198*, 2025.

Ning Yang, Hengyu Zhong, Wentao Wang, Baoliang Tian, Haijun Zhang, and Jun Wang. Linearard: Linear-memory attention distillation for rope restoration, 2026. URL https://arxiv.org/abs/2604.00004.

Longlong Zhai, Jiao Tian, Feng Yan, Lei Su, Yanjun Qin, Shaochen Jiang, Chong Peng, and Panpan Zheng. Step: Stable gradient projection for continual learning. In *Proceedings of the 2026 International Conference on Multimedia Retrieval*, ICMR '26, pp. 1404–1412, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400726170. doi: 10.1145/3805622.3810725. URL https://doi.org/10.1145/3805622.3810725.

Junhao Zheng, Chengming Shi, Xidi Cai, Qiuke Li, Duzhen Zhang, Chenxing Li, Dong Yu, and Qianli Ma. Lifelong learning of large language model based agents: A roadmap, 2026a. URL https://arxiv.org/abs/2501.07278.

Yaoyue Zheng, Yin Zhang, Joost van de Weijer, Gido M van de Ven, Shaoyi Du, Xuetao Zhang, and Zhiqiang Tian. Revisiting weight regularization for low-rank continual learning. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=pZj2DhfaVD.

Hao Zhu, Yifei Zhang, Junhao Dong, and Piotr Koniusz. Bilora: Almost-orthogonal parameter spaces for continual learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25613–25622, 2025. doi: 10.1109/CVPR52734.2025.02385.

# A SUPPLEMENTARY METHOD DETAILS

This appendix provides supplementary details for the method introduced in Section 3. We first summarize the notation used throughout the paper, followed by complete theoretical derivations and optimization pseudocode.

## A.1 NOTATION SUMMARY

For completeness, Tab. 4 summarizes the principal notation used throughout the Method and its theoretical analysis.

Table 4: Summary of the principal notation used in the proposed method.

| Notation | Definition |
|---|---|
| *Problem Setup and LoRA Decomposition* | |
| $\mathcal{T}_t$ | The $t$-th continual-learning task. |
| $T,\ t$ | Total number of tasks and task index, $t \in \{1, \dots, T\}$. |
| $p$ | Adapted attention projection, $p \in \{k, v\}$. |
| $d,\ r$ | Projection dimension and LoRA rank, with $r \ll d$. |
| $\mathbf{W}_0^p,\ \mathbf{W}_t^p$ | Frozen pre-trained and task-$t$ adapted weights of projection $p$. |
| $\Delta\mathbf{W}_{\mathrm{sh},t}^p$ | Shared low-rank update for projection $p$ during task $t$. |
| $\Delta\mathbf{W}_{\mathrm{sp},t}^p$ | Task-specific low-rank update for projection $p$ during task $t$. |
| $\mathbf{A}_{\mathrm{sh},t}^p,\ \mathbf{B}_{\mathrm{sh},t}^p$ | Shared LoRA factors satisfying $\Delta\mathbf{W}_{\mathrm{sh},t}^p = \mathbf{B}_{\mathrm{sh},t}^p\mathbf{A}_{\mathrm{sh},t}^p$. |
| $\mathbf{A}_{\mathrm{sp},t}^p,\ \mathbf{B}_{\mathrm{sp},t}^p$ | Task-specific LoRA factors satisfying $\Delta\mathbf{W}_{\mathrm{sp},t}^p = \mathbf{B}_{\mathrm{sp},t}^p\mathbf{A}_{\mathrm{sp},t}^p$. |
| $\mathbf{\Theta}_{\mathrm{sh},t},\ \mathbf{\Theta}_{\mathrm{sp},t}$ | Collections of shared and task-specific LoRA factors for task $t$. |
| *Retention-Controlled Knowledge Consolidation (RKC-EMA)* | |
| $i,\ K_t$ | Optimization-step index and total optimization steps for task $t$. |
| $\mathbf{\Theta}_{\mathrm{sh},t,i}$ | Online shared LoRA state after step $i$ of task $t$. |
| $\bar{\mathbf{\Theta}}_{\mathrm{sh},t,i}$ | Consolidated shared LoRA state after step $i$ of task $t$. |
| $\alpha_t$ | Local EMA consolidation coefficient for task $t$. |
| $R_t$ | Realized cumulative retention, $R_t = \alpha_t^{K_t}$. |
| $R$ | Prescribed task-level cumulative retention target. |
| $\omega_{t,i}(R)$ | Weight assigned to the online shared state at step $i$. |
| $\delta_t$ | Trajectory-radius bound used in the consolidated-drift analysis. |
| $\boldsymbol{\theta}_{\mathrm{sh},t,i},\ \bar{\boldsymbol{\theta}}_{\mathrm{sh},t,i}$ | An individual online shared factor and its consolidated counterpart. |
| *Structured Shared–Specific Knowledge Decomposition (SKD)* | |
| $\mathcal{S}_{\mathrm{sh},t}$ | Collection of online shared LoRA factors used by SKD. |
| $\mathcal{S}_{\mathrm{sp},t}^B$ | Collection of task-specific output factors $\mathbf{B}_{\mathrm{sp},t}^p$. |
| $\lambda_{\mathrm{cmp}},\ \lambda_{\mathrm{fid}},\ \lambda_{\mathrm{sp}}$ | Weights for shared compactness, historical fidelity, and task-specific sparsity. |
| $\mathcal{L}_{\mathrm{CL},t}$ | Continual-learning objective for task $t$. |
| $\mathcal{L}_{\mathrm{SKD},t}$ | Structured shared–specific regularization objective for task $t$. |
| *Norms* | |
| $\|\cdot\|_F,\ \|\cdot\|_1,\ \|\cdot\|_*$ | Frobenius, element-wise $\ell_1$, and nuclear norms, respectively. |

## A.2 COMPLETE THEORETICAL DERIVATIONS

### A.2.1 PROOF OF THEOREM 1

For an arbitrary shared factor $\boldsymbol{\theta}_{\mathrm{sh},t,i} \in \mathbf{\Theta}_{\mathrm{sh},t,i}$, the consolidation recurrence is

$$\bar{\boldsymbol{\theta}}_{\mathrm{sh},t,i} = \alpha_t\bar{\boldsymbol{\theta}}_{\mathrm{sh},t,i-1} + (1-\alpha_t)\boldsymbol{\theta}_{\mathrm{sh},t,i}. \tag{7}$$

Unrolling Eq. 7,

$$\begin{aligned}
\bar{\boldsymbol{\theta}}_{\mathrm{sh},t,K_t} &= \alpha_t\bar{\boldsymbol{\theta}}_{\mathrm{sh},t,K_t-1} + (1-\alpha_t)\boldsymbol{\theta}_{\mathrm{sh},t,K_t} \\
&= \alpha_t^2\bar{\boldsymbol{\theta}}_{\mathrm{sh},t,K_t-2} + (1-\alpha_t)\left(\alpha_t\boldsymbol{\theta}_{\mathrm{sh},t,K_t-1} + \boldsymbol{\theta}_{\mathrm{sh},t,K_t}\right) \\
&\ \ \vdots \\
&= \alpha_t^{K_t}\bar{\boldsymbol{\theta}}_{\mathrm{sh},t,0} + (1-\alpha_t)\sum_{i=1}^{K_t}\alpha_t^{K_t-i}\boldsymbol{\theta}_{\mathrm{sh},t,i}.
\end{aligned} \tag{8}$$

The realized cumulative retention is therefore

$$R_t \triangleq \alpha_t^{K_t}. \tag{9}$$

For a prescribed target $R \in (0,1)$,

$$\begin{aligned} R_t = R &\iff \alpha_t^{K_t} = R \\ &\iff K_t \log \alpha_t = \log R \\ &\iff \alpha_t = \exp\left(\frac{\log R}{K_t}\right) = R^{1/K_t}. \end{aligned} \tag{10}$$

Moreover,

$$\frac{\partial \alpha^{K_t}}{\partial \alpha} = K_t \alpha^{K_t - 1} > 0, \qquad \alpha \in (0,1), \quad K_t \geq 1, \tag{11}$$

which establishes uniqueness.

For arbitrary tasks $a$ and $b$,

$$\begin{aligned} \alpha_a^{K_a} &= \left(R^{1/K_a}\right)^{K_a} = R, \\ \alpha_b^{K_b} &= \left(R^{1/K_b}\right)^{K_b} = R, \\ \therefore \qquad \alpha_a^{K_a} &= \alpha_b^{K_b} = R, \qquad \forall K_a, K_b \geq 1. \end{aligned} \tag{12}$$

By contrast, for a fixed local coefficient $\alpha \in (0,1)$,

$$R_{\text{fixed}}(K) = \alpha^K, \qquad \frac{\partial R_{\text{fixed}}(K)}{\partial K} = \alpha^K \log \alpha < 0. \tag{13}$$

For the consolidated-drift bound, Eq. 8 gives

$$\begin{aligned} &\bar{\boldsymbol{\theta}}_{\text{sh},t,K_t} - \bar{\boldsymbol{\theta}}_{\text{sh},t,0} \\ &= \left(\alpha_t^{K_t} - 1\right) \bar{\boldsymbol{\theta}}_{\text{sh},t,0} + (1-\alpha_t) \sum_{i=1}^{K_t} \alpha_t^{K_t - i} \boldsymbol{\theta}_{\text{sh},t,i}. \end{aligned} \tag{14}$$

Using

$$\begin{aligned} (1-\alpha_t) \sum_{i=1}^{K_t} \alpha_t^{K_t - i} &= (1-\alpha_t) \sum_{j=0}^{K_t - 1} \alpha_t^j \\ &= 1 - \alpha_t^{K_t} = 1 - R, \end{aligned} \tag{15}$$

Eq. 14 becomes

$$\begin{aligned} &\bar{\boldsymbol{\theta}}_{\text{sh},t,K_t} - \bar{\boldsymbol{\theta}}_{\text{sh},t,0} \\ &= (1-\alpha_t) \sum_{i=1}^{K_t} \alpha_t^{K_t - i} \left(\boldsymbol{\theta}_{\text{sh},t,i} - \bar{\boldsymbol{\theta}}_{\text{sh},t,0}\right). \end{aligned} \tag{16}$$

Assume

$$\left\|\boldsymbol{\theta}_{\text{sh},t,i} - \bar{\boldsymbol{\theta}}_{\text{sh},t,0}\right\|_F \leq \delta_t, \qquad \forall i \in \{1, \ldots, K_t\}. \tag{17}$$

Then

$$\begin{aligned} &\left\|\bar{\boldsymbol{\theta}}_{\text{sh},t,K_t} - \bar{\boldsymbol{\theta}}_{\text{sh},t,0}\right\|_F \\ &\leq (1-\alpha_t) \sum_{i=1}^{K_t} \alpha_t^{K_t - i} \left\|\boldsymbol{\theta}_{\text{sh},t,i} - \bar{\boldsymbol{\theta}}_{\text{sh},t,0}\right\|_F \\ &\leq (1-\alpha_t) \sum_{i=1}^{K_t} \alpha_t^{K_t - i} \delta_t \\ &= (1-R)\delta_t. \end{aligned} \tag{18}$$

Hence,

$$\alpha_t = R^{1/K_t}, \qquad R_t = R, \qquad \left\|\bar{\boldsymbol{\theta}}_{\text{sh},t,K_t} - \bar{\boldsymbol{\theta}}_{\text{sh},t,0}\right\|_F \leq (1-R)\delta_t. \tag{19}$$

### A.2.2 Structural Derivations for SKD

**Shared-Factor Complexity.** For $p \in \{k, v\}$,

$$\Delta \mathbf{W}_{\mathrm{sh}}^{p} = \mathbf{B}_{\mathrm{sh}}^{p} \mathbf{A}_{\mathrm{sh}}^{p}, \qquad \mathbf{B}_{\mathrm{sh}}^{p} \in \mathbb{R}^{d \times r}, \quad \mathbf{A}_{\mathrm{sh}}^{p} \in \mathbb{R}^{r \times d}. \tag{20}$$

The factorization satisfies

$$\begin{aligned} \operatorname{rank}\left(\Delta \mathbf{W}_{\mathrm{sh}}^{p}\right) &= \operatorname{rank}\left(\mathbf{B}_{\mathrm{sh}}^{p} \mathbf{A}_{\mathrm{sh}}^{p}\right) \\ &\leq \min\left\{\operatorname{rank}(\mathbf{B}_{\mathrm{sh}}^{p}), \operatorname{rank}(\mathbf{A}_{\mathrm{sh}}^{p})\right\} \\ &\leq r. \end{aligned} \tag{21}$$

For the nuclear norm,

$$\begin{aligned} \left\|\Delta \mathbf{W}_{\mathrm{sh}}^{p}\right\|_{*} &= \left\|\mathbf{B}_{\mathrm{sh}}^{p} \mathbf{A}_{\mathrm{sh}}^{p}\right\|_{*} \\ &\leq \left\|\mathbf{B}_{\mathrm{sh}}^{p}\right\|_{F} \left\|\mathbf{A}_{\mathrm{sh}}^{p}\right\|_{F} \\ &\leq \frac{1}{4}\left(\left\|\mathbf{B}_{\mathrm{sh}}^{p}\right\|_{F} + \left\|\mathbf{A}_{\mathrm{sh}}^{p}\right\|_{F}\right)^{2}. \end{aligned} \tag{22}$$

The last step follows from

$$4xy \leq (x+y)^{2}, \qquad x, y \geq 0. \tag{23}$$

Thus,

$$\left(\left\|\mathbf{A}_{\mathrm{sh}}^{p}\right\|_{F} + \left\|\mathbf{B}_{\mathrm{sh}}^{p}\right\|_{F}\right) \downarrow \quad \Longrightarrow \quad \mathrm{UB}\left(\left\|\Delta \mathbf{W}_{\mathrm{sh}}^{p}\right\|_{*}\right) \downarrow, \tag{24}$$

where $\mathrm{UB}(\cdot)$ denotes the upper bound in Eq. 22.

**Task-Specific Selective Activation.** Let

$$b = [\mathbf{B}_{\mathrm{sp},t}^{p}]_{ij}, \qquad \phi(b) = \mathcal{L}_{\mathrm{CL}}(b) + \lambda_{\mathrm{sp}}|b|. \tag{25}$$

The first-order stationarity condition is

$$0 \in \partial\phi(b) = \frac{\partial \mathcal{L}_{\mathrm{CL}}}{\partial b} + \lambda_{\mathrm{sp}} \partial|b|. \tag{26}$$

Since

$$\partial|b| = \begin{cases} \{1\}, & b > 0, \\ [-1, 1], & b = 0, \\ \{-1\}, & b < 0, \end{cases} \tag{27}$$

at $b = 0$,

$$\begin{aligned} 0 \in \partial\phi(0) &\iff 0 \in \frac{\partial \mathcal{L}_{\mathrm{CL}}}{\partial b} + \lambda_{\mathrm{sp}}[-1, 1] \\ &\iff -\lambda_{\mathrm{sp}} \leq \frac{\partial \mathcal{L}_{\mathrm{CL}}}{\partial b} \leq \lambda_{\mathrm{sp}} \\ &\iff \left|\frac{\partial \mathcal{L}_{\mathrm{CL}}}{\partial b}\right| \leq \lambda_{\mathrm{sp}}. \end{aligned} \tag{28}$$

Equivalently,

$$\left|\frac{\partial \mathcal{L}_{\mathrm{CL}}}{\partial [\mathbf{B}_{\mathrm{sp},t}^{p}]_{ij}}\right| \leq \lambda_{\mathrm{sp}} \quad \Longrightarrow \quad [\mathbf{B}_{\mathrm{sp},t}^{p}]_{ij} = 0 \text{ is stationary.} \tag{29}$$

**Historical Fidelity.** For a shared factor $\boldsymbol{\theta}_{\mathrm{sh}}$ and its consolidated counterpart $\bar{\boldsymbol{\theta}}_{\mathrm{sh}}$,

$$\mathcal{R}_{\mathrm{fid}} = \lambda_{\mathrm{fid}} \left\|\boldsymbol{\theta}_{\mathrm{sh}} - \bar{\boldsymbol{\theta}}_{\mathrm{sh}}\right\|_{F}^{2}. \tag{30}$$

Its gradient is

$$\nabla_{\boldsymbol{\theta}_{\mathrm{sh}}} \mathcal{R}_{\mathrm{fid}} = 2\lambda_{\mathrm{fid}} \left(\boldsymbol{\theta}_{\mathrm{sh}} - \bar{\boldsymbol{\theta}}_{\mathrm{sh}}\right). \tag{31}$$

**Algorithm 1** Retention-Controlled Shared-State Consolidation

**Require:** Target retention $R$, task horizon $K_t$, initial online shared state $\mathbf{\Theta}_{\mathrm{sh},t,0}$
1: $\alpha_t \leftarrow R^{1/K_t}$
2: $\bar{\mathbf{\Theta}}_{\mathrm{sh},t,0} \leftarrow \mathbf{\Theta}_{\mathrm{sh},t,0}$
3: **for** $i = 1, \ldots, K_t$ **do**
4: Obtain updated online shared factors $\mathbf{\Theta}_{\mathrm{sh},t,i}$
5: $\bar{\mathbf{\Theta}}_{\mathrm{sh},t,i} \leftarrow \alpha_t \bar{\mathbf{\Theta}}_{\mathrm{sh},t,i-1} + (1-\alpha_t)\mathbf{\Theta}_{\mathrm{sh},t,i}$
6: **end for**
7: **return** $\bar{\mathbf{\Theta}}_{\mathrm{sh},t,K_t}$

For an isolated gradient step of size $\eta$,

$$\begin{aligned}\boldsymbol{\theta}_{\mathrm{sh}}^{+} &= \boldsymbol{\theta}_{\mathrm{sh}} - \eta \nabla_{\boldsymbol{\theta}_{\mathrm{sh}}} \mathcal{R}_{\mathrm{fid}} \\ &= \boldsymbol{\theta}_{\mathrm{sh}} - 2\eta\lambda_{\mathrm{fid}} \left(\boldsymbol{\theta}_{\mathrm{sh}} - \bar{\boldsymbol{\theta}}_{\mathrm{sh}}\right), \end{aligned} \tag{32}$$

and therefore

$$\boldsymbol{\theta}_{\mathrm{sh}}^{+} - \bar{\boldsymbol{\theta}}_{\mathrm{sh}} = (1 - 2\eta\lambda_{\mathrm{fid}}) \left(\boldsymbol{\theta}_{\mathrm{sh}} - \bar{\boldsymbol{\theta}}_{\mathrm{sh}}\right), \tag{33}$$

$$\left\|\boldsymbol{\theta}_{\mathrm{sh}}^{+} - \bar{\boldsymbol{\theta}}_{\mathrm{sh}}\right\|_F = |1 - 2\eta\lambda_{\mathrm{fid}}| \left\|\boldsymbol{\theta}_{\mathrm{sh}} - \bar{\boldsymbol{\theta}}_{\mathrm{sh}}\right\|_F. \tag{34}$$

If

$$0 < \eta\lambda_{\mathrm{fid}} < 1, \tag{35}$$

then

$$|1 - 2\eta\lambda_{\mathrm{fid}}| < 1, \tag{36}$$

and hence

$$\left\|\boldsymbol{\theta}_{\mathrm{sh}}^{+} - \bar{\boldsymbol{\theta}}_{\mathrm{sh}}\right\|_F < \left\|\boldsymbol{\theta}_{\mathrm{sh}} - \bar{\boldsymbol{\theta}}_{\mathrm{sh}}\right\|_F. \tag{37}$$

The contraction in Eq. 37 applies to the isolated fidelity step and does not imply contraction of the complete optimization objective.

**Combined Structural Effect.** The three structural components are summarized by

$$\begin{aligned} \mathcal{L}_{\mathrm{cmp}}: &\quad \|\mathbf{A}_{\mathrm{sh}}^{p}\|_F + \|\mathbf{B}_{\mathrm{sh}}^{p}\|_F \downarrow \implies \mathrm{UB}\left(\|\Delta\mathbf{W}_{\mathrm{sh}}^{p}\|_*\right) \downarrow, \\ \mathcal{L}_{\mathrm{fid}}: &\quad -\nabla\mathcal{R}_{\mathrm{fid}} \propto \bar{\boldsymbol{\theta}}_{\mathrm{sh}} - \boldsymbol{\theta}_{\mathrm{sh}}, \\ \mathcal{L}_{\mathrm{sp}}: &\quad \left|\frac{\partial\mathcal{L}_{\mathrm{CL}}}{\partial[\mathbf{B}_{\mathrm{sp},t}^{p}]_{ij}}\right| \le \lambda_{\mathrm{sp}} \implies [\mathbf{B}_{\mathrm{sp},t}^{p}]_{ij} = 0 \text{ is stationary.} \end{aligned} \tag{38}$$

### A.3 Optimization Algorithms and Pseudocode

### A.4 Evaluation Metrics

Let $A_{t,i}$ denote the accuracy on task $i$ after training on task $t$, with $t, i \in \{1, \ldots, T\}$, and let $T$ denote the total number of tasks.

**Final accuracy ($\bar{A}$).**

$$\bar{A}_T = \frac{1}{T}\sum_{i=1}^{T} A_{T,i}.$$

**Average accuracy (Avg.).**

$$\mathrm{Avg.} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{t}\sum_{i=1}^{t} A_{t,i}.$$

**Forgetting ($\bar{F}$).** At stage $t$, the average relative forgetting over previously learned tasks is

$$\bar{F}_t = \frac{1}{t}\sum_{i=1}^{t-1}\frac{\max_{s<t} A_{s,i} - A_{t,i}}{\max_{s<t} A_{s,i}}.$$

**Algorithm 2** Structured Shared–Specific Online Adaptation

**Require:** Task $\mathcal{T}_t$, online shared factors $\mathbf{\Theta}_{\mathrm{sh},t}$, current task-specific factors $\mathbf{\Theta}_{\mathrm{sp},t}$, consolidated shared state $\bar{\mathbf{\Theta}}_{\mathrm{sh},t}$

**for** each mini-batch in $\mathcal{T}_t$ **do**
  Compute $\mathcal{L}_{\mathrm{CL}}$
  $\mathcal{L}_{\mathrm{cmp}} \leftarrow \dfrac{1}{|\mathcal{S}_{\mathrm{sh}}|} \displaystyle\sum_{\boldsymbol{\theta} \in \mathcal{S}_{\mathrm{sh}}} \|\boldsymbol{\theta}\|_F$
  $\mathcal{L}_{\mathrm{sp}} \leftarrow \dfrac{1}{|\mathcal{S}^B_{\mathrm{sp},t}|} \displaystyle\sum_{\mathbf{B} \in \mathcal{S}^B_{\mathrm{sp},t}} \|\mathbf{B}\|_1$
  $\mathcal{L}_{\mathrm{fid}} \leftarrow \dfrac{1}{|\mathcal{S}_{\mathrm{sh}}|} \displaystyle\sum_{\boldsymbol{\theta} \in \mathcal{S}_{\mathrm{sh}}} \|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|_F^2$
  $\mathcal{L}_t \leftarrow \mathcal{L}_{\mathrm{CL}} + \lambda_{\mathrm{cmp}}\mathcal{L}_{\mathrm{cmp}} + \lambda_{\mathrm{sp}}\mathcal{L}_{\mathrm{sp}} + \lambda_{\mathrm{fid}}\mathcal{L}_{\mathrm{fid}}$
  Update $\mathbf{\Theta}_{\mathrm{sh},t}$ and $\mathbf{\Theta}_{\mathrm{sp},t}$ using $\mathcal{L}_t$
**end for**

**Algorithm 3** Unified Retaining-and-Structuring Procedure

**Require:** Tasks $\{\mathcal{T}_t\}_{t=1}^T$, target cumulative retention $R$

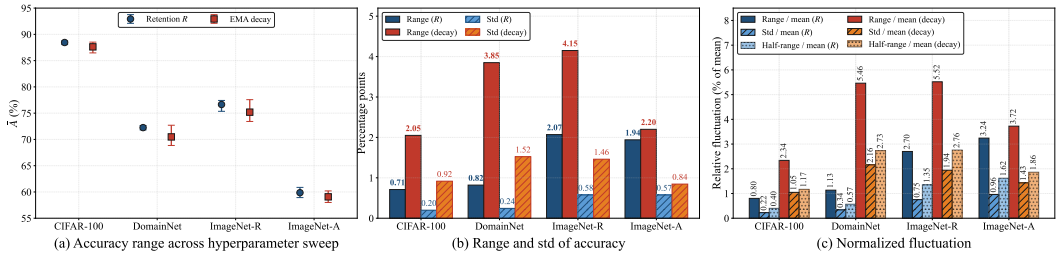
Initialize $\mathbf{\Theta}_{\mathrm{sh},1,0}$
**for** $t = 1, \ldots, T$ **do**
  Initialize current task-specific factors $\mathbf{\Theta}_{\mathrm{sp},t,0}$
  $K_t \leftarrow |\mathcal{B}_t| E_t$
  **if** $t = 1$ **then**
    **for** $i = 1, \ldots, K_1$ **do**
      Update $\mathbf{\Theta}_{\mathrm{sh},1,i}$ and $\mathbf{\Theta}_{\mathrm{sp},1,i}$ using the task-1 objective
    **end for**
    $\bar{\mathbf{\Theta}}_{\mathrm{sh},1,K_1} \leftarrow \mathbf{\Theta}_{\mathrm{sh},1,K_1}$
  **else**
    Freeze the pre-trained backbone and $\{\mathbf{\Theta}_{\mathrm{sp},j}\}_{j<t}$
    $\mathbf{\Theta}_{\mathrm{sh},t,0} \leftarrow \bar{\mathbf{\Theta}}_{\mathrm{sh},t-1,K_{t-1}}$
    $\bar{\mathbf{\Theta}}_{\mathrm{sh},t,0} \leftarrow \mathbf{\Theta}_{\mathrm{sh},t,0}$
    $\alpha_t \leftarrow R^{1/K_t}$
    **for** $i = 1, \ldots, K_t$ **do**
      Compute $\mathcal{L}_{\mathrm{CL}}$, $\mathcal{L}_{\mathrm{cmp}}$, $\mathcal{L}_{\mathrm{sp}}$, and $\mathcal{L}_{\mathrm{fid}}$
      $\mathcal{L}_t \leftarrow \mathcal{L}_{\mathrm{CL}} + \lambda_{\mathrm{cmp}}\mathcal{L}_{\mathrm{cmp}} + \lambda_{\mathrm{sp}}\mathcal{L}_{\mathrm{sp}} + \lambda_{\mathrm{fid}}\mathcal{L}_{\mathrm{fid}}$
      Update $\mathbf{\Theta}_{\mathrm{sh},t,i}$ and $\mathbf{\Theta}_{\mathrm{sp},t,i}$ using $\mathcal{L}_t$
      $\bar{\mathbf{\Theta}}_{\mathrm{sh},t,i} \leftarrow \alpha_t \bar{\mathbf{\Theta}}_{\mathrm{sh},t,i-1} + (1-\alpha_t)\mathbf{\Theta}_{\mathrm{sh},t,i}$
    **end for**
    $\mathbf{\Theta}_{\mathrm{sh},t,K_t} \leftarrow \bar{\mathbf{\Theta}}_{\mathrm{sh},t,K_t}$
  **end if**
  Freeze $\mathbf{\Theta}_{\mathrm{sp},t}$
**end for**
**return** $\bar{\mathbf{\Theta}}_{\mathrm{sh},T,K_T}$ and $\{\mathbf{\Theta}_{\mathrm{sp},t}\}_{t=1}^T$

**Stability ($S$).** The stability at stage $t$ is defined as

$$S_t = 1 - \bar{F}_t.$$

### A.5 Detailed Sensitivity Analysis of Retention Ratio and Decay Factor

This section provides a detailed sensitivity analysis of the two hyperparameters introduced by RKC-EMA: the retention ratio $R$ and the EMA decay factor. Both parameterizations are evaluated on the same set of continual-learning runs, and the results are summarized in Fig. 5 and Tables 5–6. For each sweep, the bottom block of the corresponding table reports the mean, standard deviation, maximum, minimum, and range of the accuracy across the tested values.




(a) Accuracy range across hyperparameter sweep (b) Range and std of accuracy (c) Normalized fluctuation

Figure 5: **Sensitivity of RKC-EMA to hyperparameter parameterization.** (a) Accuracy range across the swept values of the retention ratio $R$ (blue) and the EMA decay factor (red), shown as mean $\pm$ half-range. (b) Corresponding range and standard deviation of accuracy in percentage points. (c) Normalized relative fluctuation obtained by dividing each sweep by its mean accuracy. Retention-ratio control yields consistently smaller fluctuation than direct decay tuning, most notably on ImageNet-R (2.07 vs. 4.15 points).

Table 5: Sensitivity to the retention ratio $R$ on four benchmarks (%). $R$ ranges from 0.40 to 0.90; the bottom rows report the mean, standard deviation, maximum, minimum, and range across the tested values. **Bold** and *italic* denote the best and worst results, respectively. SplitLoRA (Qiu et al., 2025) is the baseline. The random seed is 1993.

| Tasks | CIFAR-100 | | DomainNet | | ImageNet-R | | ImageNet-A | |
|---|---|---|---|---|---|---|---|---|
| Methods | $\bar{A}_{10}$ (↑) | Avg. (↑) | $\bar{A}_{5}$ (↑) | Avg. (↑) | $\bar{A}_{10}$ (↑) | Avg. (↑) | $\bar{A}_{10}$ (↑) | Avg. (↑) |
| SplitLoRA | 87.99 | 92.23 | 71.30 | 77.79 | 72.05 | 75.83 | 58.87 | 64.38 |
| 0.40 | *88.05* | 92.00 | 71.82 | 78.57 | 77.10 | **80.65** | 60.47 | **70.93** |
| 0.45 | 88.48 | 92.09 | 72.21 | 78.74 | 77.08 | 80.51 | 59.67 | 70.38 |
| 0.50 | 88.65 | 92.07 | 72.02 | 78.69 | 76.85 | 80.36 | **60.87** | 70.81 |
| 0.55 | 88.23 | *91.77* | *72.01* | 78.69 | **77.42** | 80.46 | 60.00 | 70.45 |
| 0.60 | 88.39 | **92.18** | 72.28 | 78.77 | 76.97 | 80.43 | 59.87 | 70.05 |
| 0.65 | 88.33 | 92.02 | 72.45 | 78.84 | 76.95 | 80.19 | 60.00 | 70.29 |
| 0.70 | 88.37 | 91.91 | 72.63 | **78.86** | 76.28 | 79.87 | 60.40 | 70.85 |
| 0.75 | 88.56 | 92.17 | **72.64** | 78.83 | 76.95 | 80.19 | *58.93* | 69.66 |
| 0.80 | 88.66 | 92.06 | 72.32 | 78.66 | 76.25 | 79.51 | 59.87 | 69.75 |
| 0.85 | 88.56 | 92.08 | 72.32 | 78.62 | 76.05 | 79.33 | 59.47 | 69.53 |
| 0.90 | **88.76** | 92.16 | 72.14 | *78.55* | *75.35* | *78.88* | *58.93* | *69.38* |

**Retention ratio is stable over a wide range.** As shown in Fig. 5(a), RKC-EMA is insensitive to the precise value of $R$. Across $R \in [0.40, 0.90]$, the accuracy standard deviation stays below 0.58 and the performance range is at most 2.07 points on all four benchmarks. On CIFAR-100 and DomainNet, the average accuracy varies by less than 0.5 points across the entire sweep. This stability indicates that a single fixed retention target can be transferred across tasks and datasets without per-dataset tuning, which is the intended practical advantage of parameterizing consolidation at the trajectory level rather than at the update level.

**Normalized comparison.** To compare the two parameterizations on a common scale that is independent of the absolute accuracy level of each benchmark, we normalize each sweep by its mean accuracy, as shown in Fig. 5(c). The normalized fluctuation of the retention-ratio sweep remains below 2.4% across all four datasets, while that of the decay sweep exceeds 5% on ImageNet-R and DomainNet—roughly a two-fold gap. This scale-free view confirms that the advantage of retention-based control is not an artifact of the accuracy scale of individual benchmarks.

**Summary.** Overall, the results in this section show that specifying a cumulative retention target provides a more stable and transferable parameterization of shared knowledge consolidation than directly selecting the EMA decay factor. A single fixed $R$ achieves near-optimal performance across all four benchmarks, whereas the optimal decay factor varies substantially across datasets.

Table 6: Sensitivity to the EMA decay factor on four benchmarks (%). The decay factor ranges from 0.9 to 0.99999; the bottom rows report the mean, standard deviation, maximum, minimum, and range across the tested values. **Bold** and *italic* denote the best and worst results, respectively. SplitLoRA (Qiu et al., 2025) is the baseline. Seed 1993.

| Tasks | CIFAR-100 | | DomainNet | | ImageNet-R | | ImageNet-A | |
|---|---|---|---|---|---|---|---|---|
| Methods | $\bar{A}_{10}$ (↑) | Avg. (↑) | $\bar{A}_{5}$ (↑) | Avg. (↑) | $\bar{A}_{10}$ (↑) | Avg. (↑) | $\bar{A}_{10}$ (↑) | Avg. (↑) |
| SplitLoRA | 87.99 | 92.23 | 71.30 | 77.79 | 72.05 | 75.83 | 58.87 | 64.38 |
| 0.9 | 86.54 | *91.04* | 69.02 | 76.73 | 75.83 | 79.95 | **60.20** | 70.36 |
| 0.99 | *86.47* | 91.10 | *68.85* | *76.67* | **77.57** | **80.76** | 59.93 | **70.50** |
| 0.999 | **88.52** | **92.10** | 70.07 | 77.44 | 75.15 | 78.83 | *58.00* | 68.80 |
| 0.9999 | 88.33 | 91.96 | **72.70** | **78.83** | 73.98 | 78.09 | 58.73 | 68.69 |
| 0.99999 | 88.25 | 91.95 | 71.78 | 78.27 | *73.42* | *77.75* | *58.53* | *68.59* |

**Direct EMA decay tuning is substantially more sensitive.** In contrast, directly varying the EMA decay factor over $[0.9, 0.99999]$ produces much larger variations, with ranges reaching $4.15$ points and standard deviations up to $1.52$, most notably on DomainNet and ImageNet-R. This gap is expected from the parameterization itself. The decay factor controls the per-update contribution of the current shared state, so its effect accumulates exponentially with the optimization horizon; as a result, the same decay value can yield very different effective retention across tasks with different horizons. RKC-EMA instead derives the local coefficient from a prescribed cumulative retention target, which keeps the trajectory-level retention invariant to the horizon. This difference also explains the dataset-dependent optima observed in Table 6: the best decay is $0.999$ on CIFAR-100, $0.9999$ on DomainNet and ImageNet-A, and $0.99$ on ImageNet-R, whereas a single $R = 0.5$ is near-optimal across all four benchmarks (Tab. 5).

Table 7: Effect of QKV placement on top of RKC-EMA. "Params" denotes the ratio of trainable LoRA parameters to the total number of adapted projection parameters ($d \times d$ per projection). Numbers are final accuracy $\bar{A}$ (%). The best result is in **bold**.

| Strategy | Trainable ratio (%) | CIFAR-100 | | ImageNet-R | | ImageNet-A | |
|---|---|---|---|---|---|---|---|
| | | $\bar{A}_{10}$ (↑) | Avg. (↑) | $\bar{A}_{10}$ (↑) | Avg. (↑) | $\bar{A}_{10}$ (↑) | Avg. (↑) |
| Q only | 0.1867% | 87.34 | 91.37 | 74.23 | 77.44 | 59.13 | 68.31 |
| K only | 0.1867% | 87.37 | 91.60 | 74.30 | 76.96 | 57.60 | 66.78 |
| V only | 0.1867% | 87.22 | 91.78 | 76.08 | 79.56 | 59.47 | 70.37 |
| QK | 0.3652% | 87.27 | 90.80 | 74.53 | 77.86 | 59.20 | 69.03 |
| QV | 0.3652% | 87.38 | 90.76 | 76.52 | 80.16 | 59.73 | 69.88 |
| KV (Ours) | 0.3652% | 87.52 | 91.00 | 76.78 | 80.41 | 59.67 | 69.90 |
| QKV | 0.5431% | 87.40 | 91.73 | 76.93 | 80.63 | 59.87 | 70.29 |

### A.6 The analysis of the fine-tune strategy

We further study which attention projections should be adapted when RKC-EMA is applied. Following the LoRA-based PECL setup, we consider seven combinations of the query ($Q$), key ($K$), and value ($V$) projections, and report final accuracy on CIFAR-100, ImageNet-R, and ImageNet-A in Tab. 7. All runs use the same RSK configuration, and the trainable-parameter ratio is computed as the number of trainable LoRA parameters divided by the total number of adapted projection parameters ($d \times d$ per projection).

### A.7 Multi-Seed Evaluation

The main paper reports results under a single random seed, while Tables 8, 9, and 10 provide results under three random seeds (1993, 1997, and 2026) to further assess the robustness of the proposed components. Across the three seeds, RKC-EMA consistently improves the overall performance of the underlying PECL methods, although individual metrics may vary across datasets. The gains are particularly evident on CIFAR-100, DomainNet, and ImageNet-R, and are observed for different

Table 8: Comparison results on CIFAR-100, DomainNet, ImageNet-R, and ImageNet-A (in %). **Bold** and underline indicate the highest and second-highest scores, respectively. The random seed is 1993. "+ RSK" denotes combining the base method with both RKC-EMA and SKD.

| Tasks | CIFAR-100 | | DomainNet | | ImageNet-R | | ImageNet-A | |
|---|---|---|---|---|---|---|---|---|
| Methods | $\bar{A}_{10}$ (↑) | Avg. (↑) | $\bar{A}_{5}$ (↑) | Avg. (↑) | $\bar{A}_{10}$ (↑) | Avg. (↑) | $\bar{A}_{10}$ (↑) | Avg. (↑) |
| InfLoRA | 86.13 | 91.54 | 72.64 | 79.14 | 75.65 | 80.05 | 50.67 | 64.82 |
| O-LoRA | 87.44 | 91.43 | 65.56 | 75.47 | 75.87 | 79.38 | 58.27 | 68.76 |
| SD-LoRA | 87.97 | 92.14 | 71.87 | 78.39 | 70.65 | 75.71 | 56.80 | 69.43 |
| + RKC-EMA | 89.01 | 92.86 | 73.22 | 79.57 | 72.10 | 77.63 | 57.40 | 69.74 |
| + RSK | 89.08 | 92.82 | 73.37 | 79.66 | 73.37 | 77.51 | 59.20 | 70.32 |
| SplitLoRA | 87.99 | 92.23 | 71.30 | 77.79 | 72.05 | 75.83 | 58.87 | 64.38 |
| + RKC-EMA | 88.65 | 92.07 | 72.02 | 78.69 | 76.85 | 80.36 | 60.87 | 70.81 |
| + RSK | 88.34 | 92.09 | 72.74 | 79.00 | **77.35** | **80.72** | **61.20** | **71.18** |
| EWC-LoRA | 88.04 | 92.05 | 73.37 | 79.53 | 72.05 | 77.31 | 58.33 | 69.30 |
| + RKC-EMA | **89.44** | **92.98** | 73.48 | 79.87 | 70.75 | 77.72 | 59.73 | 69.74 |
| + RSK | 89.40 | 92.96 | **73.99** | **80.07** | 70.53 | 77.75 | 60.27 | 69.84 |
| Vanilla LoRA | 84.98 | 89.93 | 68.49 | 76.42 | 74.95 | 79.58 | 59.00 | 69.24 |
| + RKC-EMA | 87.52 | 91.00 | 72.15 | 78.77 | 76.78 | 80.41 | 59.67 | 69.90 |
| + RSK | 87.58 | 91.00 | 73.28 | 79.46 | 76.88 | 80.64 | 60.27 | 69.93 |

PECL backbones, including SD-LoRA, SplitLoRA, EWC-LoRA, and Vanilla LoRA. For example, RKC-EMA improves the CIFAR-100 $\bar{A}_{10}$ of EWC-LoRA by $1.40$, $2.34$, and $1.80$ points under seeds 1993, 1997, and 2026, respectively. This indicates that the benefit of retention-controlled consolidation is largely preserved across random initializations and underlying PECL methods. Adding RSK generally provides further gains over RKC-EMA, particularly on DomainNet and ImageNet-R. For instance, RSK improves the DomainNet $\bar{A}_{5}$ of EWC-LoRA + RKC-EMA from $73.48$ to $73.99$ under seed 1993, from $73.30$ to $73.79$ under seed 1997, and from $73.60$ to $73.93$ under seed 2026. Similar improvements are observed for SplitLoRA and Vanilla LoRA. Overall, the three-seed results preserve the qualitative trends of the main experiments and support the robustness and complementarity of RKC-EMA and RSK across different continual-learning settings.

Table 9: Comparison results on CIFAR-100, DomainNet, ImageNet-R, and ImageNet-A (in %). **Bold** and underline indicate the highest and second-highest scores, respectively. The random seed is 1997. "+ RSK" denotes combining the base method with both RKC-EMA and SKD.

| Tasks | CIFAR-100 | | DomainNet | | ImageNet-R | | ImageNet-A | |
|---|---|---|---|---|---|---|---|---|
| Methods | $\bar{A}_{10}$ (↑) | Avg. (↑) | $\bar{A}_{5}$ (↑) | Avg. (↑) | $\bar{A}_{10}$ (↑) | Avg. (↑) | $\bar{A}_{10}$ (↑) | Avg. (↑) |
| InfLoRA | 86.19 | 91.60 | 72.70 | 79.20 | 75.71 | 80.11 | 50.73 | 64.88 |
| O-LoRA | 87.50 | 91.49 | 65.62 | 75.53 | 75.93 | 79.44 | 58.33 | 68.82 |
| SD-LoRA | 88.03 | 92.20 | 71.93 | 78.45 | 70.71 | 75.77 | 56.86 | 69.49 |
| + RKC-EMA | 89.07 | 92.92 | 73.28 | 79.63 | 72.16 | 77.69 | 57.46 | 69.80 |
| + RSK | 89.14 | 92.88 | 73.43 | 79.72 | 73.43 | 77.57 | 59.26 | 70.38 |
| SplitLoRA | 88.05 | 92.29 | 71.36 | 77.85 | 72.11 | 75.89 | 58.93 | 64.44 |
| + RKC-EMA | 88.71 | 92.13 | 72.08 | 78.75 | 76.91 | 80.42 | 60.93 | 70.87 |
| + RSK | 88.40 | 92.15 | 72.80 | 79.06 | **77.41** | **80.78** | **61.26** | **71.24** |
| EWC-LoRA | 88.10 | 92.11 | 73.43 | 79.59 | 72.11 | 77.37 | 58.39 | 69.36 |
| + RKC-EMA | **89.58** | **93.07** | 73.55 | 79.91 | 70.81 | 77.78 | 59.79 | 69.80 |
| + RSK | 89.51 | 92.99 | **74.03** | **80.11** | 70.59 | 77.81 | 60.33 | 69.90 |
| Vanilla LoRA | 85.04 | 89.99 | 68.55 | 76.48 | 75.01 | 79.64 | 59.06 | 69.30 |
| + RKC-EMA | 87.58 | 91.06 | 72.21 | 78.83 | 76.84 | 80.47 | 59.73 | 69.96 |
| + RSK | 87.64 | 91.06 | 73.34 | 79.52 | 76.94 | 80.70 | 60.33 | 69.99 |

Table 10: Comparison results on CIFAR-100, DomainNet, ImageNet-R, and ImageNet-A (in %). **Bold** and underline indicate the highest and second-highest scores, respectively. The random seed is 2026. "+ RSK" denotes combining the base method with both RKC-EMA and SKD.

| Tasks | CIFAR-100 | | DomainNet | | ImageNet-R | | ImageNet-A | |
|---|---|---|---|---|---|---|---|---|
| Methods | $\bar{A}_{10}$ (↑) | Avg. (↑) | $\bar{A}_{5}$ (↑) | Avg. (↑) | $\bar{A}_{10}$ (↑) | Avg. (↑) | $\bar{A}_{10}$ (↑) | Avg. (↑) |
| InfLoRA | 86.38 | 91.32 | 72.64 | 79.07 | 75.27 | 79.94 | 50.13 | 64.05 |
| O-LoRA | 87.39 | 91.75 | 68.53 | 76.46 | 76.28 | 79.57 | 57.47 | 67.69 |
| SD-LoRA | 87.97 | 92.04 | 71.77 | 78.17 | 69.77 | 75.34 | 57.80 | 68.65 |
| + RKC-EMA | 89.12 | 92.96 | 73.53 | 79.69 | 72.37 | 77.52 | 57.47 | 69.19 |
| + RSK | 89.15 | 92.96 | 73.52 | 79.72 | 72.65 | 77.59 | 57.20 | 69.23 |
| SplitLoRA | 88.00 | 92.46 | 70.46 | 76.94 | 72.13 | 75.57 | 57.13 | 67.29 |
| + RKC-EMA | 88.56 | 92.27 | 72.28 | 78.82 | 76.98 | 80.30 | **60.67** | 70.17 |
| + RSK | 88.80 | 92.38 | 72.54 | 79.04 | **77.53** | **80.61** | **60.67** | 70.66 |
| EWC-LoRA | 87.95 | 92.29 | 73.11 | 79.32 | 72.97 | 77.74 | 59.60 | 69.66 |
| + RKC-EMA | **89.75** | 93.12 | 73.60 | 79.83 | 70.78 | 76.87 | 59.53 | 70.74 |
| + RSK | 89.60 | **93.16** | **73.93** | **79.96** | 70.75 | 76.96 | 59.80 | **70.76** |
| Vanilla LoRA | 84.98 | 90.95 | 68.80 | 76.63 | 75.95 | 79.77 | 57.47 | 68.23 |
| + RKC-EMA | 87.95 | 92.58 | 72.17 | 78.82 | 76.87 | 80.12 | 59.00 | 69.32 |
| + RSK | 88.06 | 92.65 | 73.25 | 79.36 | 77.02 | 80.32 | 59.40 | 69.34 |