STABLE–PLASTIC SUBSPACE CONSTRUCTION FOR DIFFERENTIALLY PRIVATE CONTINUAL LORA

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual learning (CL) adapts pretrained models to a sequence of tasks while retaining knowledge acquired from earlier tasks. When task data are sensitive, differential privacy (DP) provides record-level protection, but maintaining both adaptation and retention under privacy constraints remains challenging. We study this problem in continual low-rank adaptation (LoRA) and observe that: 1) private training increases forgetting primarily during subsequent adaptation rather than initial task acquisition; 2) jointly perturbing both LoRA factors introduces a non-negligible bilinear noise interaction; and 3) DP aggravates the stability–plasticity dilemma, hindering both knowledge retention and adaptation to new tasks. Based on these observations, we propose DP-SPA, which constructs a stable subspace from previously privatized updates and a plastic subspace from pretrained weights within its orthogonal complement. The method fixes these bases and privately optimizes only the LoRA factor, eliminating the bilinear noise interaction within each step. Before merging each task update, it rescales stable-direction coefficients using historical and current update energies. We establish a privacy guarantee and show that recalibration does not increase historical overlap and tightens a conditional upper bound on old-task loss increase. Experiments across five datasets, eight settings, and multiple privacy budgets show that DP-SPA improves average accuracy and reduces forgetting compared with the evaluated baselines.

1 INTRODUCTION

Continual learning (CL, Rolnick et al. (2019)) enables pretrained models to acquire new task knowledge sequentially while preserving previously learned capabilities. In many practical settings, however, continual adaptation is performed on sensitive task-specific data, exposing learned models to privacy attacks that can recover information about training data (Fu et al., 2024; Zhang et al., 2025; Chen et al., 2026). Differential privacy (DP) addresses this concern by limiting how much the output distribution can change when a single record is added or removed (Dwork, 2006). Maintaining effective adaptation and knowledge retention under this constraint remains challenging.

Parameter-efficient fine-tuning (PEFT) reduces the number of trainable parameters for continual adaptation (Liu et al., 2026b; Qiao & Mahdavi, 2024). Among PEFT methods, Low-Rank Adaptation (LoRA, Hu et al. (2022)) is a promising approach for CL by parameterizing each task update with two low-rank factors while keeping the pretrained backbone frozen. Existing continual LoRA methods, however, are largely designed for non-private training and often exploit task-dependent features, gradient information, or other data-derived statistics (Lu et al., 2026; He et al., 2026; Qiu et al., 2026; Liang & Li, 2024; Wu et al., 2025b). Under DP, these data-dependent operations require appropriate privacy treatment and accounting. Meanwhile, fixed-factor LoRA has been explored in private federated learning (Sun et al., 2024), where fixing one factor avoids bilinear noise interaction. However, the magnitude of this interaction in DP continual learning remains insufficiently characterized, and fixed-factor optimization alone does not specify how to select adaptation directions that balance retention and adaptation. This raises a critical question: *how should low-rank adaptation be structured to preserve historical knowledge while learning new tasks under DP?*

To answer this question, we first examine how DP affects continual LoRA and identify three key observations: 1) DP substantially exacerbates forgetting, with much of the degradation arising during subsequent private adaptation rather than initial task acquisition; 2) jointly privatizing the two LoRA

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.

factors introduces a non-negligible bilinear noise interaction; 3) DP worsens the stability–plasticity trade-off, leading to poorer knowledge retention at comparable levels of new-task adaptation.

These observations motivate DP-SPA, a **D**ifferentially **P**rivate **S**table-**P**lastic low-rank **A**daptation framework for CL. Building on fixed-factor optimization, DP-SPA addresses how to construct and update the adaptation basis across tasks under privacy constraints. It reconstructs a stable subspace from previously privatized updates and a plastic subspace from public pretrained weights within the stable subspace's orthogonal complement. The resulting bases evolve across tasks but remain fixed within each task, while DP-SGD optimizes only the corresponding LoRA up-projection matrices. To reuse historical directions while controlling their modification, DP-SPA recalibrates stable-direction coefficients using historical and current update energies before merging. Basis construction and recalibration use only public information and privatized outputs, incurring no additional privacy cost. Theoretical analysis and extensive experiments further support the effectiveness of DP-SPA.

Overall, our contributions are threefold:

- We characterize the effects of DP training on continual LoRA by separating task acquisition from forgetting during subsequent adaptation, quantifying the bilinear noise interaction under joint-factor optimization, and comparing knowledge retention at matched levels of new-task accuracy.
- We propose DP-SPA, which constructs complementary stable and plastic subspaces from privatized history and public pretrained weights, fixes them within each task for single-factor private optimization, and recalibrates stable-direction updates using historical and current energies.
- We establish the privacy guarantee of DP-SPA, characterize its stable–plastic geometry, and prove that recalibration reduces historical overlap and tightens a conditional bound on old-task loss increase. Extensive experiments across eight continual learning settings, five datasets, and multiple privacy budgets demonstrate higher average accuracy and lower forgetting than baselines.

2 Related Work

Low-Rank Adaptation for Continual Learning. Recent work has explored LoRA for CL by controlling low-rank updates to mitigate task interference. Representative approaches exploit gradient or parameter subspaces (Liang & Li, 2024; Liu & Chang, 2025; Wang et al., 2025), gradient-space splitting (Qiu et al., 2026), residual-gradient adaptation (Luo et al., 2026), magnitude–direction decoupling (Wu et al., 2025b), shared and task-specific adapters (He et al., 2025), and almost-orthogonal parameter spaces (Zhu et al., 2025). Continual LoRA merging (Qiao & Mahdavi, 2026) further consolidates sequential adaptations into a single adapter. These methods primarily address non-private CL, while components accessing private data require appropriate privacy treatment and accounting under DP. DP-SPA constructs stable and plastic subspaces from privatized updates and public pretrained weights, respectively, without accessing current private data to select bases.

Differentially Private Parameter-Efficient Fine-Tuning. Differentially private fine-tuning enables pre-trained models to adapt to sensitive downstream data with formal privacy guarantees. Parameter-efficient methods improve the privacy–utility trade-off by restricting private optimization to a small subset of parameters (Yu et al., 2021), with subsequent work exploring bias-only adaptation (Bu et al., 2024), low-rank adaptation for diffusion models (Tsai et al., 2025), and private optimization within low-dimensional gradient subspaces (Zheng et al., 2026). In federated settings, FFA-LoRA, FedASK, and LA-LoRA explore different strategies for private LoRA optimization, including fixed-factor (Sun et al., 2024), sketch-based (Wen et al., 2025), and alternating updates (Liu et al., 2026a), respectively. These studies primarily consider individual downstream tasks rather than continual adaptation. Building on fixed-factor optimization, DP-SPA balances retention and adaptation through history-aware basis construction and update recalibration under DP.

3 Understanding DP Low-Rank Continual Learning

3.1 Problem Formulation

We consider a sequence of T tasks with private datasets $\{\mathcal{D}^t\}_{t=1}^{T}$, where $\mathcal{D}^t = \{(x_i^t, y_i^t)\}_{i=1}^{N^t}$ and $y_i^t \in \mathcal{C}^t$. The class spaces are disjoint across tasks, i.e., $\mathcal{C}^t \cap \mathcal{C}^{t'} = \emptyset$ for $t \neq t'$. In the class-incremental setting, after learning task t, the model predicts over all classes seen so far, $\mathcal{C}^{1:t} =$

$\bigcup_{\tau=1}^{t} \mathcal{C}^\tau$, without access to task identity. Given a pretrained weight matrix $\boldsymbol{W}^0 \in \mathbb{R}^{m\times n}$, LoRA (Hu et al., 2022) represents the task-t low-rank adaptation by $\boldsymbol{B}^t\boldsymbol{A}^t$, where $\boldsymbol{A}^t \in \mathbb{R}^{r\times n}$ and $\boldsymbol{B}^t \in \mathbb{R}^{m\times r}$ with $r \ll \min(m,n)$. In continual adaptation, each task t starts from the model obtained after learning the preceding tasks and introduces an additive low-rank update: $\boldsymbol{W}^t = \boldsymbol{W}^{t-1} + \Delta\boldsymbol{W}^t, \Delta\boldsymbol{W}^t = \boldsymbol{B}^t\boldsymbol{A}^t$. During task t, the previously learned weights $\boldsymbol{W}^{t-1}$ are kept frozen, while both low-rank factors $\boldsymbol{A}^t$ and $\boldsymbol{B}^t$ are trainable. We initialize $\boldsymbol{A}^t$ from a Gaussian distribution and set $\boldsymbol{B}^t = \boldsymbol{0}$. We optimize the trainable LoRA parameters using DP-SGD (Abadi et al., 2016). For each sample $z_i^t = (x_i^t, y_i^t) \in \mathcal{B}^t$, where $\mathcal{B}^t \subseteq \mathcal{D}^t$ denotes a mini-batch, the per-sample gradient is clipped as $\boldsymbol{g}_i^t = \nabla_{\boldsymbol{\phi}^t}\ell(\boldsymbol{\phi}^t; z_i^t)/\max(1, \|\nabla_{\boldsymbol{\phi}^t}\ell(\boldsymbol{\phi}^t; z_i^t)\|_2/C)$, where $\boldsymbol{\phi}^t$ denotes the trainable LoRA parameters at task t and ℓ denotes the per-sample loss. The privatized mini-batch gradient is $\widetilde{\boldsymbol{g}}^t = ((\sum_{i\in\mathcal{B}^t} \boldsymbol{g}_i^t + \mathcal{N}(\boldsymbol{0}, C^2\sigma^2\boldsymbol{I}))/|\mathcal{B}^t|)$, where C is the clipping norm and σ is the Gaussian noise multiplier calibrated to the target privacy budget. Each protected record belongs to exactly one task dataset, and each task satisfies (ϵ, δ)-DP. Theorem 1 formally states the overall privacy guarantee.

Definition 1 (**(ϵ, δ)-DP**) *Let $\epsilon > 0$ and $0 < \delta < 1$. A random mechanism $\mathcal{M} : \mathcal{X}^n \to \mathcal{O}$ satisfies (ϵ, δ)-DP if, for any two neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$ that differ by the addition or removal of one record, and for any measurable $\mathcal{U} \subseteq \mathcal{O}$, $\Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{U}] \leq e^\epsilon \cdot \Pr[\mathcal{M}(\mathcal{D}') \in \mathcal{U}] + \delta$.*

3.2 Challenges of DP Low-Rank Continual Learning

We consider a 20-task class-incremental sequence on ImageNet-A (Hendrycks et al., 2021b) and fine-tune a ViT-B/16 model (Dosovitskiy et al., 2021). By default, both LoRA factors $\boldsymbol{A}$ and $\boldsymbol{B}$ are jointly updated, with DP noise added to the gradients of both factors.

Observation 1. *Differential privacy substantially exacerbates forgetting in low-rank continual adaptation, with the degradation arising mainly during subsequent adaptation rather than initial task acquisition.* Let $a_{t,k}$ denote the accuracy on task k after learning task t. We measure forgetting as $F_k^t = \max_{\tau\in\{k,\ldots,t-1\}} a_{\tau,k} - a_{t,k}$ and $F^t = \frac{1}{t-1}\sum_{k=1}^{t-1} F_k^t$. As shown in Fig. 1(a), DP ($\epsilon = 3$) incurs substantially larger forgetting than Non-DP throughout the task sequence. To examine how this degradation evolves, we fix the first $K = 5$ tasks and track their forgetting after u subsequent adaptations, $\bar{F}^u = \frac{1}{K}\sum_{k=1}^{K} F_k^{k+u}$. Fig. 1(b) shows that forgetting increases under both settings, while remaining consistently higher under DP. Fig. 1(c) further compares acquisition accuracy $a_{k,k}$ with retention accuracy $a_{k+15,k}$. Although acquisition is comparable between DP and Non-DP, retention falls to 44.20% under DP versus 51.17% under Non-DP after 15 subsequent adaptations, indicating that the additional forgetting emerges mainly during later adaptation.

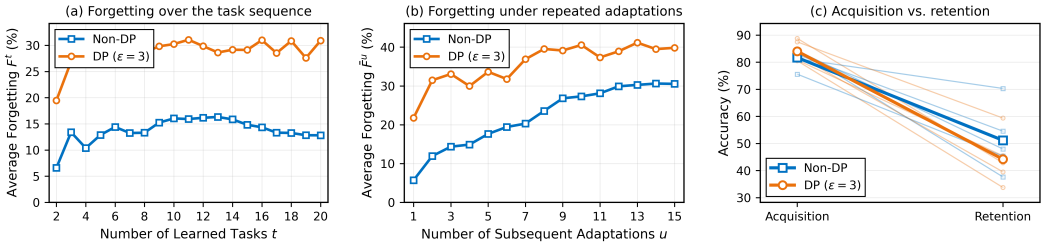

Figure 1: DP exacerbates forgetting during continual adaptation. (a) Average forgetting across tasks. (b) Early-task forgetting over subsequent adaptations. (c) Acquisition vs. retention after 15 subsequent adaptations. Light lines denote individual tasks; bold lines denote their mean.

Observation 2. *Jointly privatizing the two LoRA factors introduces a non-negligible bilinear noise interaction during sequential adaptation.* Consider one DP-SGD update within a task. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ denote the LoRA factors after the corresponding noise-free clipped-gradient update, and let $\boldsymbol{\xi}_A$ and $\boldsymbol{\xi}_B$ denote the parameter perturbations induced by the Gaussian gradient noise in this update. The resulting factors can be written as $\widetilde{\boldsymbol{A}} = \boldsymbol{A} + \boldsymbol{\xi}_A, \widetilde{\boldsymbol{B}} = \boldsymbol{B} + \boldsymbol{\xi}_B$. Accordingly, the privacy-induced distortion in the effective LoRA update is

$$\widetilde{\boldsymbol{B}}\widetilde{\boldsymbol{A}} - \boldsymbol{B}\boldsymbol{A} = \boldsymbol{B}\boldsymbol{\xi}_A + \boldsymbol{\xi}_B\boldsymbol{A} + \boldsymbol{\xi}_B\boldsymbol{\xi}_A. \tag{1}$$

The first two terms describe the first-order propagation of the factor-wise perturbations, whereas $\boldsymbol{\xi}_B\boldsymbol{\xi}_A$ captures their second-order interaction induced by the bilinear LoRA parameterization.

We quantify the relative magnitude of the bilinear interaction over the continual sequence under $\sigma = 0.8$. We aggregate the perturbation energies over DP-SGD steps and LoRA modules:

$$R_t = \frac{\sqrt{\sum_{\ell,q} \|\boldsymbol{\xi}^{(\ell)}_{B,t,q}\boldsymbol{\xi}^{(\ell)}_{A,t,q}\|_F^2}}{\sqrt{\sum_{\ell,q} \|\boldsymbol{B}^{(\ell)}_{t,q}\boldsymbol{\xi}^{(\ell)}_{A,t,q} + \boldsymbol{\xi}^{(\ell)}_{B,t,q}\boldsymbol{A}^{(\ell)}_{t,q}\|_F^2}}, \tag{2}$$

where q indexes mini-batch DP-SGD updates within task t and ℓ indexes LoRA modules. The bilinear interaction remains consistently non-negligible throughout the 20-task sequence, despite substantial task-wise variation (Fig. 2).

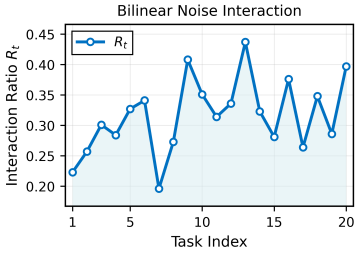

Figure 2: Bilinear interaction ratio.

Observation 3. *Differential privacy degrades the stability–plasticity trade-off, making it harder to retain previous knowledge while effectively adapting to new tasks.* We characterize plasticity P by new-task accuracy and stability by old-task accuracy. We divide the 20-task sequence into four consecutive stages of five tasks each and average the new-task and old-task accuracies within each stage to obtain the stage-wise plasticity–stability operating points. As shown in Fig. 3(a), both DP and Non-DP shift toward lower plasticity and stability as adaptation progresses, while DP consistently operates in a lower-stability region. To isolate the stability cost from differences in new-task adaptation, Fig. 3(b) compares the two settings at matched plasticity levels. The stability gap remains positive across the matched range, showing that comparable plasticity under DP is associated with systematically lower stability. Fig. 3(c) further shows that this trade-off degradation persists across privacy budgets and becomes less severe as the privacy constraint is relaxed.

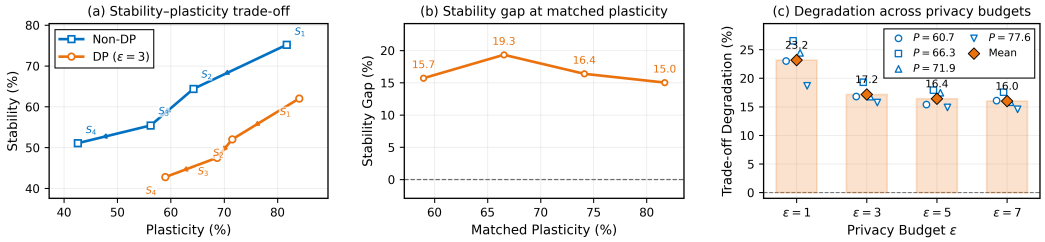

Figure 3: Stability–plasticity trade-off under differential privacy. (a) Stage-wise stability–plasticity trajectories for Non-DP and DP ($\epsilon = 3$). (b) Stability gap between Non-DP and DP at matched plasticity. (c) Trade-off degradation across privacy budgets, where markers denote stability gaps at individual matched plasticity levels P and bars denote their mean.

4 Proposed Method

Overview. To address the issues in Section 3.2, we propose DP-SPA, which constructs stable and plastic LoRA bases before accessing current-task private data and restricts direct access to these data to DP-SGD. At task t, the stable subspace $\mathcal{S}^t$ captures dominant historical adaptation directions from $\boldsymbol{H}^{t-1}$, a statistic of previously privatized model updates. The plastic subspace $\mathcal{P}^t$ is constructed from the pretrained weight $\boldsymbol{W}^0$ within the orthogonal complement of $\mathcal{S}^t$, providing complementary directions for new-task adaptation (Fig. 4(A)). The down-projection matrices $\boldsymbol{A}^t_S$ and $\boldsymbol{A}^t_P$ remain fixed, while only the corresponding up-projection matrices $\boldsymbol{B}^t_S$ and $\boldsymbol{B}^t_P$ are trained using DP-SGD, eliminating the bilinear noise interaction (Fig. 4(B)). After private optimization, we recalibrate the stable-branch update using historical update energy and current update energy computed from the DP-SGD outputs. The resulting LoRA update $\Delta\boldsymbol{W}^t$ is merged into the continual model and incorporated into $\boldsymbol{H}^t$ to guide stable-subspace construction for the next task (Fig. 4(C)).

4.1 Privacy-Constrained Stable-Plastic Subspace Construction

At task t, the LoRA adaptation is parameterized by $\boldsymbol{B}^t\boldsymbol{A}^t$. $\boldsymbol{A}^t \in \mathbb{R}^{r\times n}$ determines the low-dimensional input-side subspace through which adaptation is expressed, while $\boldsymbol{B}^t \in \mathbb{R}^{m\times r}$ maps this representation to the output space Zhu et al. (2024); Tian et al. (2024). We therefore use $\boldsymbol{A}^t$ to determine the low-rank adaptation subspace. Under continual adaptation, the limited rank budget

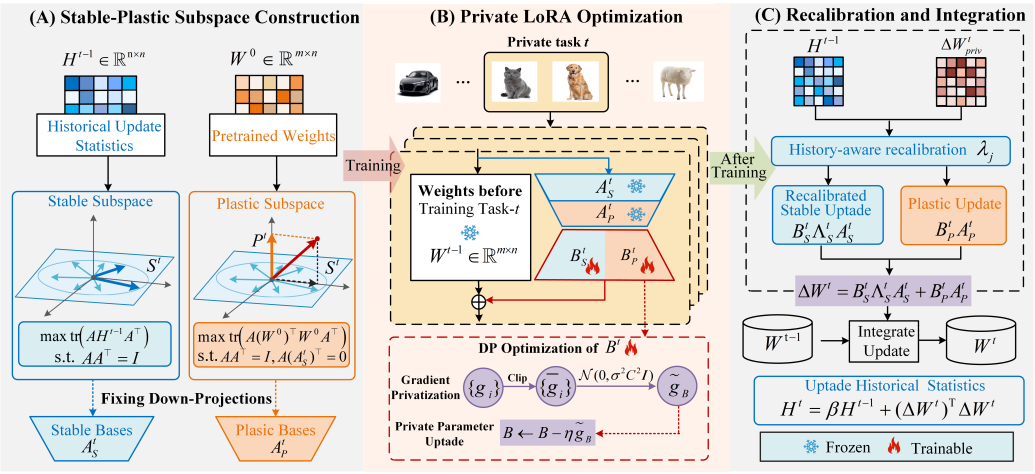

Figure 4: Overview of DP-SPA. (A) Construct the stable subspace $\mathcal{S}^t$ from privatized update history $\boldsymbol{H}^{t-1}$ and the plastic subspace $\mathcal{P}^t$ from pretrained weights $\boldsymbol{W}^0$ within $(\mathcal{S}^t)^\perp$. (B) Within task t, fix $\boldsymbol{A}_S^t, \boldsymbol{A}_P^t$ and optimize only $\boldsymbol{B}_S^t, \boldsymbol{B}_P^t$ with DP-SGD. (C) Recalibrate stable-branch updates using historical and current update energies, merge $\Delta\boldsymbol{W}^t$ into the model, and update $\boldsymbol{H}^t$ for the next task.

must preserve historically occupied directions while retaining capacity outside them for subsequent tasks. We therefore decompose the row space of $\boldsymbol{A}^t$ into orthogonal stable and plastic components:

$$\boldsymbol{A}^t = \begin{bmatrix} \boldsymbol{A}_S^t \\ \boldsymbol{A}_P^t \end{bmatrix}, \qquad r_S + r_P = r, \tag{3}$$

where the rows of $\boldsymbol{A}_S^t \in \mathbb{R}^{r_S \times n}$ and $\boldsymbol{A}_P^t \in \mathbb{R}^{r_P \times n}$ span the stable subspace $\mathcal{S}^t$ and plastic subspace $\mathcal{P}^t$. The ranks r_S and r_P determine the capacity allocated to the two components. Both bases are constructed before accessing the current private dataset and satisfy $\boldsymbol{A}_S^t(\boldsymbol{A}_S^t)^\top = \boldsymbol{I}$, $\boldsymbol{A}_P^t(\boldsymbol{A}_P^t)^\top = \boldsymbol{I}$, and $\boldsymbol{A}_P^t(\boldsymbol{A}_S^t)^\top = \boldsymbol{0}$. Accordingly, the LoRA up-projection is partitioned as $\boldsymbol{B}^t = [\boldsymbol{B}_S^t, \boldsymbol{B}_P^t]$.

Historical stable subspace. We construct the stable subspace from the updates that have already been privatized and merged into the continual model. To summarize their directional usage, we maintain the exponentially weighted second-order statistic $\boldsymbol{H}^{t-1} = \sum_{i=1}^{t-1} \beta^{t-1-i} (\Delta\boldsymbol{W}^i)^\top \Delta\boldsymbol{W}^i$, where $\beta \in (0, 1]$ controls the decay of earlier task updates. For any unit direction $\boldsymbol{a} \in \mathbb{R}^n$, $\boldsymbol{a}^\top \boldsymbol{H}^{t-1}\boldsymbol{a} = \sum_{i=1}^{t-1} \beta^{t-1-i} \left\|\Delta\boldsymbol{W}^i \boldsymbol{a}\right\|_2^2$. Hence, $\boldsymbol{H}^{t-1}$ summarizes the accumulated update energy along different right-side directions. We define the stable subspace as the r_S-dimensional subspace that preserves the largest amount of this accumulated historical update energy:

$$\boldsymbol{A}_S^t = \arg\max_{\boldsymbol{A}} \operatorname{tr}\left(\boldsymbol{A}\boldsymbol{H}^{t-1}\boldsymbol{A}^\top\right) \quad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{I}. \tag{4}$$

Here, $\boldsymbol{A} \in \mathbb{R}^{r_S \times n}$, and $\arg\max$ denotes a selected maximizer. Let $\boldsymbol{H}^{t-1} = \boldsymbol{V}_H^t \boldsymbol{\Lambda}_H^t (\boldsymbol{V}_H^t)^\top$ be its eigendecomposition, where $\boldsymbol{V}_H^t$ contains orthonormal eigenvectors and $\boldsymbol{\Lambda}_H^t = \operatorname{diag}(h_1^t, \ldots, h_n^t)$ with $h_1^t \geq \cdots \geq h_n^t \geq 0$, the solution is $\boldsymbol{A}_S^t = (\boldsymbol{V}_{H,1:r_S}^t)^\top$ by the Rayleigh–Ritz principle. Thus, $\mathcal{S}^t$ captures the dominant directions of historical adaptation.

Residual plastic subspace. To reserve adaptation capacity outside the historical stable subspace, we allocate the remaining r_P directions to its orthogonal complement. Since the basis is constructed before accessing the current private dataset, we use the pretrained weight $\boldsymbol{W}^0$ as a structural prior. Specifically, we select the subspace with maximal pretrained response energy:

$$\boldsymbol{A}_P^t = \arg\max_{\boldsymbol{A}} \operatorname{tr}\left(\boldsymbol{A}(\boldsymbol{W}^0)^\top \boldsymbol{W}^0 \boldsymbol{A}^\top\right) \quad \text{s.t.} \quad \boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{I},\ \boldsymbol{A}(\boldsymbol{A}_S^t)^\top = \boldsymbol{0}. \tag{5}$$

Here, $\boldsymbol{A} \in \mathbb{R}^{r_P \times n}$. Let $\boldsymbol{U}_\perp^t \in \mathbb{R}^{n \times (n - r_S)}$ be an orthonormal basis of $(\mathcal{S}^t)^\perp$. Setting $\boldsymbol{A} = \boldsymbol{Z}^\top (\boldsymbol{U}_\perp^t)^\top$, where $\boldsymbol{Z}^\top \boldsymbol{Z} = \boldsymbol{I}_{r_P}$, reduces Eq. 5 to maximizing $\operatorname{tr}(\boldsymbol{Z}^\top \boldsymbol{C}_P^t \boldsymbol{Z})$, where $\boldsymbol{C}_P^t = (\boldsymbol{U}_\perp^t)^\top (\boldsymbol{W}^0)^\top \boldsymbol{W}^0 \boldsymbol{U}_\perp^t$. Let $\mu_1^t \geq \cdots \geq \mu_{n-r_S}^t \geq 0$ denote the eigenvalues of $\boldsymbol{C}_P^t$. For

$r_P \le n - r_S$, let $\boldsymbol{Z}^t_{P,1:r_P}$ contain orthonormal eigenvectors associated with the r_P largest eigenvalues $\mu^t_1, \dots, \mu^t_{r_P}$ of $\boldsymbol{C}^t_P$. Then

$$\boldsymbol{A}^t_P = (\boldsymbol{Z}^t_{P,1:r_P})^\top (\boldsymbol{U}^t_\perp)^\top. \tag{6}$$

Thus, $\boldsymbol{A}^t_P(\boldsymbol{A}^t_P)^\top = \boldsymbol{I}_{r_P}$ and $\boldsymbol{A}^t_P(\boldsymbol{A}^t_S)^\top = \boldsymbol{0}$, including when $\boldsymbol{C}^t_P$ is rank deficient. The resulting plastic subspace maximizes pretrained response energy within $(\mathcal{S}^t)^\perp$.

Initialization. For task $t = 1$, we initialize the fixed basis using the dominant right-singular directions of the pretrained weight: $\boldsymbol{W}^0 = \boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^\top$, $\boldsymbol{A}^1 = \boldsymbol{V}^\top_{0,1:r}$. After task 1, $\Delta\boldsymbol{W}^1$ initializes $\boldsymbol{H}^1$, enabling the stable–plastic construction for subsequent tasks.

4.2 Private LoRA Optimization

Once the stable–plastic basis is constructed, we keep $\boldsymbol{A}^t$ fixed throughout task t and privately optimize only $\boldsymbol{B}^t$ within each LoRA module. Fixing $\boldsymbol{A}^t$ preserves the adaptation geometry constructed from historical and pretrained information and eliminates the bilinear noise interaction identified in **Observation 2**. We apply the DP-SGD mechanism defined in Section 3.1 to $\boldsymbol{B}^t$: $\boldsymbol{B}^t \leftarrow \boldsymbol{B}^t - \eta\,\widetilde{\boldsymbol{g}}^t_B$, where $\eta > 0$ is the learning rate and $\widetilde{\boldsymbol{g}}^t_B$ denotes the DP-SGD gradient with respect to $\boldsymbol{B}^t$. Following the single-step perturbation analysis in **Observation 2**, let $\boldsymbol{\xi}^t_B$ denote the Gaussian perturbation applied to $\boldsymbol{B}^t$. Because $\boldsymbol{A}^t$ is fixed, the induced distortion in the effective LoRA update reduces to $(\boldsymbol{B}^t + \boldsymbol{\xi}^t_B)\boldsymbol{A}^t - \boldsymbol{B}^t\boldsymbol{A}^t = \boldsymbol{\xi}^t_B\boldsymbol{A}^t$. Compared with joint privatization, the perturbation terms associated with the down-projection, $\boldsymbol{B}^t\boldsymbol{\xi}^t_A$ and $\boldsymbol{\xi}^t_B\boldsymbol{\xi}^t_A$, are eliminated. Hence, private perturbation propagates only through the fixed adaptation basis, without basis noise or bilinear noise interaction between the two LoRA factors.

4.3 History-Aware Recalibration and Continual Integration

After private optimization, the resulting LoRA update is decomposed as $\Delta\boldsymbol{W}^t_{\text{priv}} = \boldsymbol{B}^t_S\boldsymbol{A}^t_S + \boldsymbol{B}^t_P\boldsymbol{A}^t_P$. Stable directions differ in their historical update energy and current adaptation strength. We therefore recalibrate their contributions individually before model integration. For the j-th stable direction, we define its historical update energy as $h^t_j = \boldsymbol{a}^t_{S,j}\boldsymbol{H}^{t-1}(\boldsymbol{a}^t_{S,j})^\top$. Here, $\boldsymbol{a}^t_{S,j}$ denotes the j-th row of $\boldsymbol{A}^t_S$. Since $\boldsymbol{a}^t_{S,j}$ is an eigenvector of $\boldsymbol{H}^{t-1}$, h^t_j equals its associated eigenvalue. We measure the current adaptation strength by the energy of the corresponding rank-one update, $c^t_j = \|\boldsymbol{b}^t_{S,j}\boldsymbol{a}^t_{S,j}\|^2_F = \|\boldsymbol{b}^t_{S,j}\|^2_2$, where $\boldsymbol{b}^t_{S,j}$ is the j-th column of $\boldsymbol{B}^t_S$ and $\|\boldsymbol{a}^t_{S,j}\|_2 = 1$.

We determine a recalibration coefficient by balancing fidelity to the privately learned update against modification along historically active directions. Specifically, for each stable direction we solve $c^t_j(1-\lambda)^2 + (h^t_j + \tau)\lambda^2$ over $\lambda \in [0,1]$, where $\tau > 0$ is a small regularization constant. The first term penalizes deviation from the privately learned rank-one update, whereas the second penalizes retaining a large current update along a historically active direction. The unique minimizer is

$$\lambda^t_j = c^t_j/(h^t_j + c^t_j + \tau), \qquad 0 \le \lambda^t_j < 1. \tag{7}$$

When historical update energy dominates, the corresponding current stable-direction update is attenuated more strongly. When the current adaptation strength is comparatively larger, a greater fraction of the privately learned update is retained.

Let $\boldsymbol{\Lambda}^t_S = \mathrm{diag}(\lambda^t_1, \dots, \lambda^t_{r_S})$, the final update is then

$$\Delta\boldsymbol{W}^t = \boldsymbol{B}^t_S\boldsymbol{\Lambda}^t_S\boldsymbol{A}^t_S + \boldsymbol{B}^t_P\boldsymbol{A}^t_P. \tag{8}$$

The plastic branch is retained without recalibration.

For $t = 1$, we use $\Delta\boldsymbol{W}^1 = \boldsymbol{B}^1\boldsymbol{A}^1$ without recalibration. At every task, we merge the update via $\boldsymbol{W}^t = \boldsymbol{W}^{t-1} + \Delta\boldsymbol{W}^t$ and update the historical statistic recursively: $\boldsymbol{H}^t = \beta\boldsymbol{H}^{t-1} + (\Delta\boldsymbol{W}^t)^\top\Delta\boldsymbol{W}^t$. Hence, $\boldsymbol{H}^t$ summarizes the accumulated energy of the updates merged into the continual model and guides stable-subspace construction for task $t+1$.

5 Theoretical Analysis

We establish the privacy guarantee, characterize the stable–plastic geometry, and analyze recalibration and its conditional effect on old-task loss. Complete proofs are deferred to Appendix C.

Theorem 1 (Privacy Guarantee) *Assume public pretrained weights and a fixed public task partition, with each protected record belonging to exactly one task. Under add/remove adjacency, suppose each task mechanism is (ϵ_t, δ_t)-DP conditional on any preceding private transcript, and all remaining operations use only public information and privatized outputs. Then the full transcript of released models and statistics is $(\max_t \epsilon_t, \max_t \delta_t)$-DP. In particular, assigning the same (ϵ, δ) budget to every task gives an (ϵ, δ) guarantee for the entire sequence.*

Theorem 1 allows later tasks to depend adaptively on earlier private outputs. Since basis construction, recalibration, merging, and history updates use only public information and previously privatized outputs, they constitute post-processing and incur no additional privacy cost.

For any update $\boldsymbol{U} \in \mathbb{R}^{m\times n}$, define the historical overlap energy $\mathcal{I}_t(\boldsymbol{U}) = \operatorname{tr}\big(\boldsymbol{U}\boldsymbol{H}^{t-1}\boldsymbol{U}^\top\big) = \sum_{i<t} \beta^{t-1-i}\|\boldsymbol{U}(\Delta\boldsymbol{W}^i)^\top\|_F^2$, which measures its energy along historical update directions.

Theorem 2 (Complementary Stable–Plastic Geometry) *For $t \geq 2$, $\boldsymbol{A}^t(\boldsymbol{A}^t)^\top = \boldsymbol{I}_r$, and the following statements hold. (i) Historical energy. Over all $\boldsymbol{U} \in \mathbb{R}^{r_S\times n}$ satisfying $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}_{r_S}$, the stable basis satisfies $\max_{\boldsymbol{U}} \operatorname{tr}\big(\boldsymbol{U}\boldsymbol{H}^{t-1}\boldsymbol{U}^\top\big) = \operatorname{tr}\big(\boldsymbol{A}_S^t\boldsymbol{H}^{t-1}(\boldsymbol{A}_S^t)^\top\big) = \sum_{j=1}^{r_S} h_j^t$. Every plastic update obeys $\mathcal{I}_t(\boldsymbol{B}_P^t\boldsymbol{A}_P^t) \leq h_{r_S+1}^t\|\boldsymbol{B}_P^t\|_F^2$. (ii) Pretrained response. Over all $\boldsymbol{U} \in \mathbb{R}^{r_P\times n}$ satisfying $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}_{r_P}$ and $\boldsymbol{U}(\boldsymbol{A}_S^t)^\top = \boldsymbol{0}$, the plastic basis satisfies $\max_{\boldsymbol{U}} \|\boldsymbol{W}^0\boldsymbol{U}^\top\|_F^2 = \|\boldsymbol{W}^0(\boldsymbol{A}_P^t)^\top\|_F^2 = \sum_{j=1}^{r_P}\mu_j^t$. (iii) Additional adaptation capacity. For every target update $\boldsymbol{U} \in \mathbb{R}^{m\times n}$, $\min_{\boldsymbol{X}_S\in\mathbb{R}^{m\times r_S}} \|\boldsymbol{U} - \boldsymbol{X}_S\boldsymbol{A}_S^t\|_F^2 - \min_{\boldsymbol{X}\in\mathbb{R}^{m\times r}} \|\boldsymbol{U} - \boldsymbol{X}\boldsymbol{A}^t\|_F^2 = \|\boldsymbol{U}(\boldsymbol{A}_P^t)^\top\|_F^2$. Moreover, $\|\boldsymbol{X}\boldsymbol{A}^t\|_F = \|\boldsymbol{X}\|_F$ for every $\boldsymbol{X} \in \mathbb{R}^{m\times r}$.*

Theorem 3 (Optimal Recalibration and Historical Overlap Reduction) *Condition on the realized private coefficient matrix $\boldsymbol{B}^t$ and history $\boldsymbol{H}^{t-1}$. For stable columns with $c_j^t = 0$, impose $\boldsymbol{x}_j = \boldsymbol{0}$, where $\boldsymbol{x}_j$ denotes the j-th column of $\boldsymbol{X}_S$. Then $\widehat{\boldsymbol{B}}^t = [\boldsymbol{B}_S^t\boldsymbol{\Lambda}_S^t, \boldsymbol{B}_P^t]$ is the unique minimizer of $\min_{\boldsymbol{X}=[\boldsymbol{X}_S,\boldsymbol{X}_P]}\big\{\frac{1}{2}\|\boldsymbol{X} - \boldsymbol{B}^t\|_F^2 + \frac{1}{2}\sum_{j:c_j^t>0}\frac{h_j^t+\tau}{c_j^t}\|\boldsymbol{x}_j\|_2^2\big\}$. Furthermore, $\Delta\boldsymbol{W}^t(\boldsymbol{A}_P^t)^\top = \boldsymbol{B}_P^t$, $\|\Delta\boldsymbol{W}_{\text{priv}}^t - \Delta\boldsymbol{W}^t\|_F^2 = \sum_{j=1}^{r_S} c_j^t(1-\lambda_j^t)^2$, $\|\Delta\boldsymbol{W}_{\text{priv}}^t\|_F^2 - \|\Delta\boldsymbol{W}^t\|_F^2 = \sum_{j=1}^{r_S} c_j^t(1-(\lambda_j^t)^2)$, and $\mathcal{I}_t(\Delta\boldsymbol{W}_{\text{priv}}^t) - \mathcal{I}_t(\Delta\boldsymbol{W}^t) = \sum_{j=1}^{r_S} h_j^t c_j^t(1-(\lambda_j^t)^2) \geq 0$.*

Recalibration balances fidelity to the private update against an energy-normalized historical penalty, preserving plastic coefficients without increasing historical overlap.

Theorem 4 (Conditional Forgetting Certificate) *Let F_{old} be an old-task population loss and let $\boldsymbol{W}^{t-1}$ be the pre-merge weight. Suppose $\|\nabla F_{\text{old}}(\boldsymbol{W}^{t-1})\|_F \leq \nu$. Assume that F_{old} is twice continuously differentiable along the relevant segments and that, for $\boldsymbol{U} \in \{\Delta\boldsymbol{W}_{\text{priv}}^t, \Delta\boldsymbol{W}^t\}$, $s \in [0,1]$, and every perturbation $\boldsymbol{V} \in \mathbb{R}^{m\times n}$, $\nabla^2 F_{\text{old}}(\boldsymbol{W}^{t-1} + s\boldsymbol{U})[\boldsymbol{V},\boldsymbol{V}] \leq \kappa\operatorname{tr}(\boldsymbol{V}\boldsymbol{H}^{t-1}\boldsymbol{V}^\top) + \rho\|\boldsymbol{V}\|_F^2$, where $\nu, \kappa, \rho \geq 0$. Define $\mathcal{U}(\boldsymbol{U}) = \nu\|\boldsymbol{U}\|_F + \frac{\kappa}{2}\mathcal{I}_t(\boldsymbol{U}) + \frac{\rho}{2}\|\boldsymbol{U}\|_F^2$. Then $F_{\text{old}}(\boldsymbol{W}^{t-1} + \boldsymbol{U}) - F_{\text{old}}(\boldsymbol{W}^{t-1}) \leq \mathcal{U}(\boldsymbol{U})$ for $\boldsymbol{U} \in \{\Delta\boldsymbol{W}_{\text{priv}}^t, \Delta\boldsymbol{W}^t\}$, and $\mathcal{U}(\Delta\boldsymbol{W}_{\text{priv}}^t) - \mathcal{U}(\Delta\boldsymbol{W}^t) = \nu\big(\|\Delta\boldsymbol{W}_{\text{priv}}^t\|_F - \|\Delta\boldsymbol{W}^t\|_F\big) + \frac{1}{2}\sum_{j=1}^{r_S}(\kappa h_j^t + \rho)c_j^t(1-(\lambda_j^t)^2) \geq 0$.*

Under the stated assumptions, recalibration does not increase the upper bound on old-task loss increase, but need not reduce the actual loss. The same argument extends to multiple adapted layers under a joint curvature bound accounting for cross-layer interactions.

6 Experiments

6.1 Experimental Setups

Evaluation Benchmarks and Metrics. We evaluate DP-SPA on 5 benchmarks: ImageNet-R (Hendrycks et al., 2021a), ImageNet-A (Hendrycks et al., 2021b), CIFAR-100 (Krizhevsky et al., 2009), CUB (Wah et al., 2011) and DomainNet (Peng et al., 2019). ImageNet-R contains 30,000 images from 200 classes with diverse styles. ImageNet-A comprises 7,500 real-world adversarial images from 200 categories. CIFAR-100 contains 60,000 32×32 images from 100 classes. DomainNet covers 345 classes across 6 visual domains. Following prior works (Gao et al., 2024; He et al., 2026), we select the top 200 most populous classes for evaluation.

We follow commonly used evaluation metrics in CL (Tan et al., 2024; Zhao et al., 2024; Wu et al., 2025a): Last Accuracy ($\mathcal{A}_{Last}$), which measures the final performance over all classes, and Average Accuracy ($\mathcal{A}_{Avg}$), which averages the performance across all incremental sessions.

Implementation Details. We use ViT-B/16 (Dosovitskiy et al., 2021) pre-trained on ImageNet-21K (Russakovsky et al., 2015) as the backbone. Each task is trained for 20 epochs with a batch size of 48 using DP-SGD with cosine learning-rate decay. The initial learning rate is 0.02 for ImageNet-R, 0.005 for CUB, and 0.01 for the other datasets. We insert DP-SPA into all attention blocks with $r = 32$, allocating $r_S = r_P = 16$, and train a cosine classifier (He et al., 2026) alongside the LoRA coefficients. We select the clipping threshold from $\{0.1, 0.5, 1, 2, 5\}$ and use $C = 1$. The main experiments use $\epsilon = 3$, and we additionally evaluate $\epsilon \in \{1, 3, 5, 7\}$, with $\delta = 10^{-5}$ throughout. We use the RDP accountant implemented in Opacus (Yousefpour et al., 2021) to calibrate the Gaussian noise multiplier σ to the target (ϵ, δ)-DP guarantee. All experiments are conducted using NVIDIA RTX 5090 GPUs, and the reported results are averaged over three random seeds.

6.2 Main Results

Comparison with Baseline Methods. For a controlled comparison under DP, we consider **AB-LoRA**, which jointly updates both LoRA factors $\boldsymbol{A}$ and $\boldsymbol{B}$, and a stronger baseline, **FA-LoRA**, which fixes $\boldsymbol{A}$ and updates only $\boldsymbol{B}$. Tables 1, 2 summarize the results across different datasets and incremental settings. Our method substantially outperforms AB-LoRA across all settings. On ImageNet-R with T=10, DP-SPA increases $\mathcal{A}_{Last}$ from 56.86% to 71.85% and $\mathcal{A}_{Avg}$ from 63.83% to 79.90%, corresponding to gains of 14.99 and 16.07 percentage points, respectively. The gains increase to 22.31 and 22.22 percentage points for T=20. Compared with the stronger FA-LoRA, our method consistently achieves higher accuracy across all datasets and settings. On ImageNet-R, DP-SPA improves $\mathcal{A}_{Last}$ and $\mathcal{A}_{Avg}$ by 3.60 and 2.70 percentage points for T=20, respectively. Figure 5 shows that DP-SPA maintains higher overall and old-task accuracy while preserving comparable new-task accuracy, demonstrating a better stability–plasticity trade-off.

Table 1: Experimental results for 10 incremental tasks on ImageNet-R, ImageNet-A, CIFAR-100, and CUB. The best results are highlighted in **bold**.

Method	ImageNet-R (T=10)		ImageNet-A (T=10)		CIFAR-100 (T=10)		CUB (T=10)	
	$\mathcal{A}_{Last}\uparrow$	$\mathcal{A}_{Avg}\uparrow$	$\mathcal{A}_{Last}\uparrow$	$\mathcal{A}_{Avg}\uparrow$	$\mathcal{A}_{Last}\uparrow$	$\mathcal{A}_{Avg}\uparrow$	$\mathcal{A}_{Last}\uparrow$	$\mathcal{A}_{Avg}\uparrow$
AB-LoRA	$56.86_{\pm 1.37}$	$63.83_{\pm 1.54}$	$51.69_{\pm 1.22}$	$62.23_{\pm 0.92}$	$85.42_{\pm 0.36}$	$90.40_{\pm 0.42}$	$72.37_{\pm 0.90}$	$79.33_{\pm 0.67}$
FA-LoRA	$68.55_{\pm 0.77}$	$77.46_{\pm 0.89}$	$54.01_{\pm 0.78}$	$64.42_{\pm 0.71}$	$86.56_{\pm 0.30}$	$90.63_{\pm 0.41}$	$74.02_{\pm 0.85}$	$80.26_{\pm 0.71}$
DP-SPA	$\mathbf{71.85_{\pm 0.62}}$	$\mathbf{79.90_{\pm 0.57}}$	$\mathbf{57.21_{\pm 0.58}}$	$\mathbf{66.65_{\pm 0.63}}$	$\mathbf{87.07_{\pm 0.22}}$	$\mathbf{91.19_{\pm 0.35}}$	$\mathbf{75.78_{\pm 0.64}}$	$\mathbf{82.67_{\pm 0.59}}$

Table 2: Experimental results for 5 and 20 incremental tasks on ImageNet-R, 20 tasks on ImageNet-A, and 10 tasks on DomainNet. The best results are highlighted in **bold**.

Method	ImageNet-R (T=5)		ImageNet-R (T=20)		ImageNet-A (T=20)		DomainNet (T=10)	
	$\mathcal{A}_{Last}\uparrow$	$\mathcal{A}_{Avg}\uparrow$	$\mathcal{A}_{Last}\uparrow$	$\mathcal{A}_{Avg}\uparrow$	$\mathcal{A}_{Last}\uparrow$	$\mathcal{A}_{Avg}\uparrow$	$\mathcal{A}_{Last}\uparrow$	$\mathcal{A}_{Avg}\uparrow$
AB-LoRA	$62.48_{\pm 0.86}$	$70.73_{\pm 0.74}$	$45.83_{\pm 2.96}$	$56.12_{\pm 2.41}$	$41.44_{\pm 2.23}$	$54.11_{\pm 1.80}$	$76.53_{\pm 1.12}$	$83.22_{\pm 1.05}$
FA-LoRA	$71.66_{\pm 0.59}$	$79.20_{\pm 0.61}$	$64.54_{\pm 0.93}$	$75.64_{\pm 0.79}$	$45.25_{\pm 1.42}$	$58.12_{\pm 1.04}$	$78.07_{\pm 0.64}$	$84.83_{\pm 0.65}$
DP-SPA	$\mathbf{75.45_{\pm 0.51}}$	$\mathbf{82.06_{\pm 0.42}}$	$\mathbf{68.14_{\pm 0.75}}$	$\mathbf{78.34_{\pm 0.67}}$	$\mathbf{48.44_{\pm 0.84}}$	$\mathbf{60.89_{\pm 0.78}}$	$\mathbf{79.45_{\pm 0.59}}$	$\mathbf{85.61_{\pm 0.72}}$

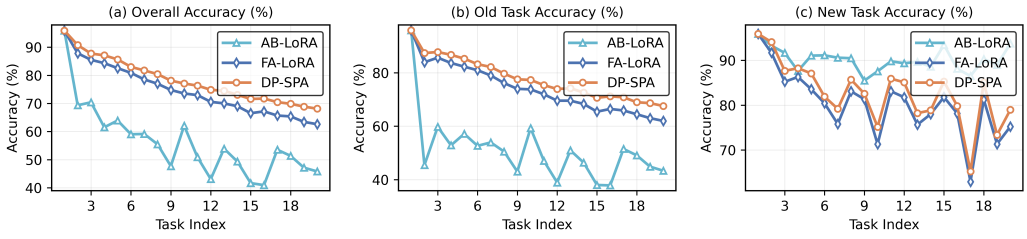

Figure 5: Overall, old task and new task accuracy on ImageNet-R across 20 incremental tasks.

Table 3: Experimental results under different ϵ on ImageNet-A and ImageNet-R with $T = 20$.

Dataset	Method	$\epsilon = 1$		$\epsilon = 3$		$\epsilon = 5$		$\epsilon = 7$	
		$\mathcal{A}_{Last} \uparrow$	$\mathcal{A}_{Avg} \uparrow$	$\mathcal{A}_{Last} \uparrow$	$\mathcal{A}_{Avg} \uparrow$	$\mathcal{A}_{Last} \uparrow$	$\mathcal{A}_{Avg} \uparrow$	$\mathcal{A}_{Last} \uparrow$	$\mathcal{A}_{Avg} \uparrow$
ImageNet-R	AB-LoRA	$3.26_{\pm 4.77}$	$11.37_{\pm 3.25}$	$45.83_{\pm 2.96}$	$56.12_{\pm 2.41}$	$48.73_{\pm 2.66}$	$57.42_{\pm 1.95}$	$50.82_{\pm 1.23}$	$60.28_{\pm 0.96}$
	FA-LoRA	$34.45_{\pm 2.52}$	$55.73_{\pm 2.70}$	$64.54_{\pm 0.93}$	$75.64_{\pm 0.79}$	$67.25_{\pm 0.88}$	$78.36_{\pm 0.73}$	$70.76_{\pm 0.61}$	$78.64_{\pm 0.84}$
	DP-SPA	$\mathbf{48.69_{\pm 1.98}}$	$\mathbf{60.44_{\pm 1.94}}$	$\mathbf{68.14_{\pm 0.75}}$	$\mathbf{78.34_{\pm 0.67}}$	$\mathbf{69.85_{\pm 0.70}}$	$\mathbf{79.65_{\pm 0.72}}$	$\mathbf{72.46_{\pm 0.55}}$	$\mathbf{80.27_{\pm 0.54}}$
ImageNet-A	AB-LoRA	$31.03_{\pm 3.35}$	$45.12_{\pm 2.87}$	$41.44_{\pm 2.23}$	$54.11_{\pm 1.80}$	$43.33_{\pm 0.66}$	$55.08_{\pm 0.73}$	$44.08_{\pm 0.68}$	$55.46_{\pm 0.49}$
	FA-LoRA	$43.30_{\pm 2.95}$	$56.07_{\pm 2.96}$	$45.25_{\pm 1.42}$	$58.12_{\pm 1.04}$	$47.61_{\pm 0.63}$	$59.75_{\pm 0.75}$	$48.24_{\pm 0.49}$	$60.07_{\pm 0.47}$
	DP-SPA	$\mathbf{45.22_{\pm 1.54}}$	$\mathbf{58.46_{\pm 1.25}}$	$\mathbf{48.44_{\pm 0.84}}$	$\mathbf{60.89_{\pm 0.78}}$	$\mathbf{49.70_{\pm 0.63}}$	$\mathbf{59.92_{\pm 0.56}}$	$\mathbf{50.43_{\pm 0.41}}$	$\mathbf{61.38_{\pm 0.44}}$

Effect of Privacy Budgets. Table 3 compares different ϵ on ImageNet-R and ImageNet-A with T=20. DP-SPA achieves the highest mean accuracy across all evaluated budgets. At $\epsilon = 1$ on ImageNet-R, AB-LoRA achieves only 3.26% in $\mathcal{A}_{Last}$, whereas FA-LoRA and DP-SPA achieve 34.45% and 48.69%, respectively. The improvement of FA-LoRA over AB-LoRA supports the benefit of fixing $\boldsymbol{A}$ under strong privacy noise. DP-SPA further outperforms FA-LoRA by 14.24 percentage points, demonstrating the additional benefit of its stable–plastic design and recalibration. As ϵ increases, this gap narrows, while DP-SPA retains its advantage on both datasets.

6.3 Ablation Studies.

Effect of Stable and Plastic bases. We ablate the stable and plastic bases to examine their individual contributions. As shown in Table 4, both $\boldsymbol{A}_S^t$-only ($(r_S, r_P) = (32, 0)$) and $\boldsymbol{A}_P^t$-only ($(0, 32)$) consistently outperform AB-LoRA, with $\boldsymbol{A}_S^t$-only performing better across all settings. This advantage becomes more pronounced for longer task sequences, highlighting the increasing importance of preserving historically important directions. Combining $\boldsymbol{A}_S^t$ and $\boldsymbol{A}_P^t$ consistently achieves the best performance, demonstrating their complementary roles in balancing stability and plasticity.

Table 4: Ablation studies on the stable and plastic bases on ImageNet-R and ImageNet-A.

Variant	**ImageNet-R** (T=10)		**ImageNet-R** (T=20)		**ImageNet-A** (T=10)		**ImageNet-A** (T=20)	
	$\mathcal{A}_{Last} \uparrow$	$\mathcal{A}_{Avg} \uparrow$	$\mathcal{A}_{Last} \uparrow$	$\mathcal{A}_{Avg} \uparrow$	$\mathcal{A}_{Last} \uparrow$	$\mathcal{A}_{Avg} \uparrow$	$\mathcal{A}_{Last} \uparrow$	$\mathcal{A}_{Avg} \uparrow$
AB-LoRA	$56.86_{\pm 1.37}$	$63.83_{\pm 1.54}$	$45.83_{\pm 2.96}$	$56.12_{\pm 2.41}$	$51.69_{\pm 1.22}$	$62.23_{\pm 0.92}$	$41.44_{\pm 2.23}$	$54.11_{\pm 1.80}$
$\boldsymbol{A}_S^t$-only	$70.53_{\pm 0.36}$	$78.05_{\pm 0.48}$	$67.76_{\pm 0.60}$	$77.15_{\pm 0.46}$	$56.17_{\pm 0.54}$	$65.13_{\pm 0.70}$	$47.05_{\pm 0.65}$	$59.43_{\pm 0.52}$
$\boldsymbol{A}_P^t$-only	$68.81_{\pm 0.52}$	$77.94_{\pm 0.57}$	$65.80_{\pm 0.74}$	$76.32_{\pm 0.81}$	$54.82_{\pm 0.61}$	$64.96_{\pm 0.52}$	$45.19_{\pm 0.90}$	$58.24_{\pm 0.68}$
$\boldsymbol{A}_S^t$+$\boldsymbol{A}_P^t$	$\mathbf{71.85_{\pm 0.62}}$	$\mathbf{79.90_{\pm 0.57}}$	$\mathbf{68.14_{\pm 0.75}}$	$\mathbf{78.34_{\pm 0.67}}$	$\mathbf{57.21_{\pm 0.58}}$	$\mathbf{66.65_{\pm 0.63}}$	$\mathbf{48.44_{\pm 0.84}}$	$\mathbf{60.89_{\pm 0.78}}$

Effect of History-Aware Recalibration. Table 5 isolates recalibration while retaining the stable–plastic subspace construction. Recalibration improves $\mathcal{A}_{Last}$ from 67.25% to 68.14% and $\mathcal{A}_{Avg}$ from 77.49% to 78.34%. This shows that recalibrating stable-direction updates further improves continual adaptation under privacy constraints.

Table 5: Effect of recalibration on ImageNet-R with $T = 20$, $r = 32$, and $\epsilon = 3$.

Variant	$\mathcal{A}_{Last} \uparrow$	$\mathcal{A}_{Avg} \uparrow$
DP-SPA w/o recal.	$67.25_{\pm 0.69}$	$77.49_{\pm 0.72}$
DP-SPA	$\mathbf{68.14_{\pm 0.75}}$	$\mathbf{78.34_{\pm 0.67}}$

7 Discussion and Conclusion

We study differentially private continual LoRA and observe increased forgetting during subsequent adaptation and lower retention at comparable new-task accuracy in the evaluated settings. Joint-factor privatization also introduces a within-step bilinear noise interaction. DP-SPA constructs a stable basis from privatized updates and a plastic basis from pretrained weights within its orthogonal complement, fixes both bases during private up-projection optimization, and recalibrates stable-direction updates using historical and current update energies. Experiments show improved average accuracy and retention across the evaluated settings and privacy budgets, supporting the reuse of privatized history without additional privacy cost. Recalibration guarantees non-increasing historical overlap; under the stated gradient and curvature assumptions, it also yields a non-increasing upper bound on old-task loss increase, without guaranteeing lower actual loss.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Differentially private bias-term fine-tuning of foundation models. In *Forty-first International Conference on Machine Learning*, 2024.

Yuetian Chen, Yuntao Du, Kaiyuan Zhang, Ashish Kundu, Charles Fleming, Bruno Ribeiro, and Ninghui Li. Window-based membership inference attacks against fine-tuned large language models. In *35th USENIX Security Symposium (USENIX Security 26)*, Baltimore, MD, August 2026. USENIX Association.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pp. 1–12. Springer, 2006.

Wenjie Fu, Huandong Wang, Chen Gao, Guanghua Liu, Yong Li, and Tao Jiang. Membership inference attacks against fine-tuned large language models via self-prompt calibration. *Advances in Neural Information Processing Systems*, 37:134981–135010, 2024.

Zhanxin Gao, Jun Cen, and Xiaobin Chang. Consistent prompting for rehearsal-free continual learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28463–28473. IEEE, 2024.

Jiangpeng He, Zhihao Duan, and Fengqing Zhu. Cl-lora: Continual low-rank adaptation for rehearsal-free class-incremental learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30534–30544. IEEE, 2025.

Lingfeng He, De Cheng, Huaijie Wang, Xi Yang, Nannan Wang, and Xinbo Gao. Task-driven subspace decomposition for knowledge sharing and isolation in lora-based continual learning. *arXiv preprint arXiv:2603.00191*, 2026.

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 *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329. IEEE, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

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

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

Jin Liu, Yinbin Miao, Ning Xi, and Junkang Liu. Rethinking lora for privacy-preserving federated learning in large models. *arXiv preprint arXiv:2602.19926*, 2026a.

Jingren Liu, Zhong Ji, YunLong Yu, Jiale Cao, Yanwei Pang, Jungong Han, and Xuelong Li. Parameter-efficient fine-tuning for continual learning: A neural tangent kernel perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026b.

Xuan Liu and Xiaobin Chang. Lora subtraction for drift-resistant space in exemplar-free continual learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15308–15318. IEEE, 2025.

Yue Lu, Shizhou Zhang, De Cheng, Yinghui Xing, Guoqiang Liang, Nannan Wang, Peng Wang, and Yanning Zhang. Continual low-rank adaptation via cumulative unified optimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Mao-Lin Luo, Zi-Hao Zhou, Yi-Lin Zhang, Yuanyu Wan, Min-Ling Zhang, and Tong Wei. Keeplora: Continual learning with residual gradient adaptation. In *International Conference on Learning Representations*, volume 2026, pp. 136017–136040, 2026.

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

Fuli Qiao and Mehrdad Mahdavi. Learn more, but bother less: parameter efficient continual learning. *Advances in Neural Information Processing Systems*, 37:97476–97498, 2024.

Fuli Qiao and Mehrdad Mahdavi. Merge before forget: A single lora continual learning via continual merging. In *International Conference on Learning Representations*, volume 2026, pp. 8133–8159, 2026.

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. In *International Conference on Learning Representations*, volume 2026, pp. 48458–48480, 2026.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving lora in privacy-preserving federated learning. In *International conference on learning representations*, volume 2024, pp. 17978–17994, 2024.

Yuwen Tan, Qinhao Zhou, Xiang Xiang, Ke Wang, Yuchuan Wu, and Yongbin Li. Semantically-shifted incremental adapter-tuning is a continual vitransformer. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23252–23262. IEEE, 2024.

Chunlin Tian, Zhan Shi, Zhijiang Guo, Li Li, and Chengzhong Xu. Hydralora: An asymmetric lora architecture for efficient fine-tuning. *Advances in Neural Information Processing Systems*, 37: 9565–9584, 2024.

Yu-Lin Tsai, Yizhe Li, Chia-Mu Yu, Xuebin Ren, Po-Yu Chen, Zekai Chen, and Francois Buet-Golfouse. Differentially private fine-tuning of diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4561–4571. IEEE, 2025.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Xiequn Wang, Zhan Zhuang, and Yu Zhang. Plan: Proactive low-rank allocation for continual learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2909–2918. IEEE, 2025.

Ming Wen, Jiaqi Zhu, Yuedong Xu, Yipeng Zhou, and DINGDING HAN. Differentially private federated low rank adaptation beyond fixed-matrix. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Fangwen Wu, Lechao Cheng, Shengeng Tang, Xiaofeng Zhu, Chaowei Fang, Dingwen Zhang, and Meng Wang. Navigating semantic drift in task-agnostic class-incremental learning. In *Forty-second International Conference on Machine Learning*, 2025a.

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. In *International Conference on Learning Representations*, volume 2025, pp. 58337–58353, 2025b.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, et al. Opacus: User-friendly differential privacy library in pytorch. *arXiv preprint arXiv:2109.12298*, 2021.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*, 2021.

Kaiyuan Zhang, Siyuan Cheng, Hanxi Guo, Yuetian Chen, Zian Su, Shengwei An, Yuntao Du, Charles Fleming, Ashish Kundu, Xiangyu Zhang, et al. {SOFT}: Selective data obfuscation for protecting {LLM} fine-tuning against membership inference attacks. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 8135–8154, 2025.

Linglan Zhao, Xuerui Zhang, Ke Yan, Shouhong Ding, and Weiran Huang. Safe: Slow and fast parameter-efficient tuning for continual learning with pre-trained models. *Advances in Neural Information Processing Systems*, 37:113772–113796, 2024.

Lele Zheng, Xiang Wang, Tao Zhang, Yang Cao, Ke Cheng, and Yulong Shen. Differentially private subspace fine-tuning for large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28830–28838, 2026.

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. IEEE, 2025.

Jiacheng Zhu, Kristjan Greenewald, Kimia Nadjahi, Haitz Sáez de Ocáriz Borde, Rickard Brüel Gabrielsson, Leshem Choshen, Marzyeh Ghassemi, Mikhail Yurochkin, and Justin Solomon. Asymmetry in low-rank adapters of foundation models. In *Forty-first International Conference on Machine Learning*, 2024.

8 ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. In this study, no human subjects or animal experimentation was involved. All datasets used, including **ImageNet-R, ImageNet-A, CIFAR-100, CUB, DomainNet**, were sourced in compliance with relevant usage guidelines, ensuring no violation of privacy. We have taken care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

9 REPRODUCIBILITY STATEMENT

We have made every effort to ensure the reproducibility of the results presented in this paper. Detailed experimental settings, including training procedures, model configurations, hyperparameters, and hardware specifications, are provided in Section 6.1 and Appendix E.1. The source code and scripts required to reproduce our experiments will be made publicly available upon acceptance.

10 AI USE STATEMENT

Large language models (LLMs) were used to assist with manuscript writing and polishing, including sentence rephrasing, grammar checking, improving readability and clarity, and checking the consistency of theoretical arguments and mathematical notation.

The LLMs were not used to originate the research ideas or design the methods or experiments. The research and analyses were conducted by the authors. The authors reviewed and revised all AI-assisted content to ensure its accuracy and consistency with the underlying research. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

APPENDIX CONTENTS

A Table of Notations 15

B DP-SPA Algorithm 15

C Proofs of the Main Theoretical Results 15

C.1 Privacy guarantee 16

C.2 Stable–plastic geometry 17

C.3 Recalibration guarantee 19

C.4 Conditional old-task loss bound 21

D Supplementary Analysis of the Bilinear Noise Interaction 23

E Supplementary Experimental Information 23

E.1 Experimental Details 23

E.2 Additional Results 24

A TABLE OF NOTATIONS

Table 6 summarizes the main notations used throughout the paper.

Table 6: Summary of the main notations.

Notation	Description
$T,\ t$	Number of tasks and current task index
$\mathcal{D}^t,\ \mathcal{C}^t$	Private dataset and class set of task t
$\boldsymbol{W}^0 \in \mathbb{R}^{m\times n}$	Pretrained weight matrix
$\boldsymbol{W}^t$	Weight matrix after completing task t
$r_S,\ r_P,\ r = r_S + r_P$	Stable, plastic, and total LoRA ranks
$\mathcal{S}^t,\ \mathcal{P}^t$	Stable and plastic input-side subspaces
$\boldsymbol{A}_S^t,\ \boldsymbol{A}_P^t$	Orthonormal row bases of $\mathcal{S}^t$ and $\mathcal{P}^t$
$\boldsymbol{A}^t = [(\boldsymbol{A}_S^t)^\top, (\boldsymbol{A}_P^t)^\top]^\top$	Fixed LoRA basis for task t
$\boldsymbol{B}^t = [\boldsymbol{B}_S^t, \boldsymbol{B}_P^t]$	Privately trained LoRA coefficient matrix
$\boldsymbol{H}^{t-1}$	Historical statistic of previously merged updates
β	Decay factor for historical updates
h_j^t	Eigenvalue h_j^t of the history matrix
c_j^t	current stable-column energy c_j^t
$\lambda_j^t,\ \boldsymbol{\Lambda}_S^t$	Direction-wise recalibration coefficient and its diagonal matrix
$\Delta\boldsymbol{W}_{\text{priv}}^t$	Private task update before recalibration
$\Delta\boldsymbol{W}^t$	Merged task update; recalibration is omitted at $t = 1$
$\mathcal{I}_t(\boldsymbol{U})$	Historical overlap energy of an update $\boldsymbol{U}$
$(\epsilon_t, \delta_t),\ (\epsilon, \delta)$	Task-specific and common record-level privacy budgets
$C,\ \sigma$	Per-record clipping norm and Gaussian noise multiplier
$\mathcal{A}_{Last},\ \mathcal{A}_{Avg}$	Last Accuracy and Average Accuracy
τ	Regularization constant in history-aware recalibration
$\boldsymbol{C}_P^t,\ \mu_j^t$	Restricted pretrained Gram matrix and its j-th eigenvalue

B DP-SPA ALGORITHM

Algorithm 1 summarizes the complete training procedure of DP-SPA, including stable–plastic basis construction, private LoRA optimization, history-aware recalibration, and model merging.

C PROOFS OF THE MAIN THEORETICAL RESULTS

The following four subsections prove the corresponding main-text theorems. The final subsection analyzes the fresh-noise interaction in Observation 2. Unless stated otherwise, geometric and recalibration results concern one adapted matrix at a fixed task $t \geq 2$, with $0 \leq r_S$, $0 \leq r_P$, and $r_S + r_P = r \leq n$. Define

$$\boldsymbol{P}^t = (\boldsymbol{A}^t)^\top \boldsymbol{A}^t, \qquad \boldsymbol{Q}_S^t = \boldsymbol{I}_n - (\boldsymbol{A}_S^t)^\top \boldsymbol{A}_S^t.$$

The history matrix is positive semidefinite, with ordered eigenvalues $h_1^t \geq \cdots \geq h_n^t \geq 0$. Its associated overlap functional is

$$\mathcal{I}_t(\boldsymbol{U}) = \operatorname{tr}(\boldsymbol{U}\boldsymbol{H}^{t-1}\boldsymbol{U}^\top) = \sum_{i<t} \beta^{t-1-i} \|\boldsymbol{U}(\Delta\boldsymbol{W}^i)^\top\|_F^2. \tag{9}$$

The second identity follows by substituting the definition of $\boldsymbol{H}^{t-1}$ and taking traces. This functional measures weighted overlap with historical updates; a connection to old-task loss requires the assumptions in Theorem 4.

Algorithm 1 Differentially Private Stable-Plastic Low-rank Adaptation (DP-SPA)

Input: Pretrained model $\boldsymbol{W}^0$, private task sequence $\{\mathcal{D}^t\}_{t=1}^T$, ranks r_S, r_P with $r = r_S + r_P$, DP parameters (C, σ), learning rate η, decay β, regularization constant $\tau > 0$
Output: Updated continual model $\boldsymbol{W}^T$ with merged LoRA updates
Initialize historical statistic $\boldsymbol{H}^0 \leftarrow \mathbf{0}$
for task $t = 1$ to T **do**
  if $t = 1$ **then**
    Initialize $\boldsymbol{A}^1$ from $\boldsymbol{W}^0$: $\boldsymbol{W}^0 = \boldsymbol{U}_0 \boldsymbol{\Sigma}_0 \boldsymbol{V}_0^\top$, $\boldsymbol{A}^1 \leftarrow \boldsymbol{V}_{0,1:r}^\top$
  else
    Construct stable basis: $\boldsymbol{H}^{t-1} = \boldsymbol{V}_H^t \boldsymbol{\Lambda}_H^t (\boldsymbol{V}_H^t)^\top$, $\boldsymbol{A}_S^t \leftarrow (\boldsymbol{V}_{H,1:r_S}^t)^\top$
    Compute $\boldsymbol{C}_P^t \leftarrow (\boldsymbol{U}_\perp^t)^\top (\boldsymbol{W}^0)^\top \boldsymbol{W}^0 \boldsymbol{U}_\perp^t$
    Construct plastic basis: $\boldsymbol{A}_P^t \leftarrow (\boldsymbol{Z}_{P,1:r_P}^t)^\top (\boldsymbol{U}_\perp^t)^\top$
    $\boldsymbol{A}^t \leftarrow [(\boldsymbol{A}_S^t)^\top, (\boldsymbol{A}_P^t)^\top]^\top$
  end if
  Initialize $\boldsymbol{B}^t \leftarrow \mathbf{0}$, freeze $\boldsymbol{A}^t$, and privately optimize $\boldsymbol{B}^t$ on $\mathcal{D}^t$ with DP-SGD
  if $t = 1$ **then**
    $\Delta \boldsymbol{W}^1 \leftarrow \boldsymbol{B}^1 \boldsymbol{A}^1$
  else
    Decompose $\boldsymbol{B}^t = [\boldsymbol{B}_S^t, \boldsymbol{B}_P^t]$
    Compute $h_j^t = \boldsymbol{a}_{S,j}^t \boldsymbol{H}^{t-1} (\boldsymbol{a}_{S,j}^t)^\top$ and $c_j^t = \|\boldsymbol{b}_{S,j}^t\|_2^2$ for $j = 1, \ldots, r_S$
    Compute $\lambda_j^t = c_j^t / (h_j^t + c_j^t + \tau)$ for $j = 1, \ldots, r_S$ and set $\boldsymbol{\Lambda}_S^t \leftarrow \mathrm{diag}(\lambda_1^t, \ldots, \lambda_{r_S}^t)$
    $\Delta \boldsymbol{W}^t \leftarrow \boldsymbol{B}_S^t \boldsymbol{\Lambda}_S^t \boldsymbol{A}_S^t + \boldsymbol{B}_P^t \boldsymbol{A}_P^t$
  end if
  Integrate the LoRA update: $\boldsymbol{W}^t \leftarrow \boldsymbol{W}^{t-1} + \Delta \boldsymbol{W}^t$
  Update the historical statistic: $\boldsymbol{H}^t \leftarrow \beta \boldsymbol{H}^{t-1} + (\Delta \boldsymbol{W}^t)^\top \Delta \boldsymbol{W}^t$
end for

C.1 Proof of the Privacy Guarantee

Proof of Theorem 1. Let $\mathcal{D} = (\mathcal{D}^1, \ldots, \mathcal{D}^T)$, $\mathcal{D}' = (\mathcal{D}'^1, \ldots, \mathcal{D}'^T)$ be neighboring continual datasets under the stated add/remove adjacency. Since each protected record belongs to exactly one fixed task partition, there is a task s such that $\mathcal{D}^s$ and $\mathcal{D}'^s$ differ by one record, while $\mathcal{D}^t = \mathcal{D}'^t$ for all $t \neq s$.

Let Z_t denote the privatized outputs of task t, including the privatized information used to construct any state passed to subsequent tasks, and write $Z_{<s} = (Z_1, \ldots, Z_{s-1})$. Under the theorem's assumptions, all remaining carried state is public or computed from these privatized outputs. Tasks preceding s access identical datasets under $\mathcal{D}$ and $\mathcal{D}'$, so $Z_{<s}$ has the same distribution under both sequences. Denote this common distribution by μ.

For a fixed prefix z, let $M_s(\mathcal{D}^s; z)$ denote the task-s mechanism. By assumption, this mechanism is (ϵ_s, δ_s)-DP with respect to $\mathcal{D}^s$ for every preceding private transcript z. For any measurable event E in the joint output space of $(Z_{<s}, Z_s)$, define $E_z = \{y : (z, y) \in E\}$. Applying the conditional DP inequality and integrating against the common prefix distribution gives

$$\begin{aligned}
&\Pr_{\mathcal{D}}\big[(Z_{<s}, Z_s) \in E\big] \\
&\quad = \int \Pr\big[M_s(\mathcal{D}^s; z) \in E_z\big]\, \mu(dz) \\
&\quad \leq \int \big(e^{\epsilon_s} \Pr\big[M_s(\mathcal{D}'^s; z) \in E_z\big] + \delta_s\big)\, \mu(dz) \\
&\quad = e^{\epsilon_s} \Pr_{\mathcal{D}'}\big[(Z_{<s}, Z_s) \in E\big] + \delta_s.
\end{aligned}$$

Thus, the joint transcript through task s satisfies the (ϵ_s, δ_s) privacy inequality for this neighboring pair.

For every subsequent task $t > s$, $\mathcal{D}^t = \mathcal{D}'^t$. The remaining computation accesses only these unchanged datasets, public information, and the preceding privatized outputs. Therefore, for the fixed

neighboring pair, it is the same randomized post-processing of $(Z_{<s}, Z_s)$ under both sequences, with the later datasets treated as fixed auxiliary inputs. The post-processing property preserves the (ϵ_s, δ_s) inequality for the complete transcript, including all released models and derived statistics. The same argument applies with $\mathcal{D}$ and $\mathcal{D}'$ interchanged.

Because the differing record may belong to any task s, the complete continual mechanism satisfies

$$\left(\max_{1\le t\le T} \epsilon_t,\ \max_{1\le t\le T} \delta_t\right)\text{-DP}.$$

In particular, assigning every task the same (ϵ, δ) budget gives an (ϵ, δ) guarantee for the entire sequence. □

Application to DP-SPA. The stable basis is constructed from $\boldsymbol{H}^{t-1}$, which depends only on previously privatized and merged updates. The plastic basis is constructed from this stable basis and the public pretrained weights $\boldsymbol{W}^0$. After private optimization, the current energies $c_j^t = \|\boldsymbol{b}_{S,j}^t\|_2^2$ are computed from the privatized coefficients $\boldsymbol{B}^t$, while the historical energies h_j^t are computed from $\boldsymbol{H}^{t-1}$. Consequently, the recalibration coefficients λ_j^t, the merged update $\Delta\boldsymbol{W}^t$, the model $\boldsymbol{W}^t$, and the updated history $\boldsymbol{H}^t$ are all functions of public information and privatized outputs. These operations therefore incur no additional privacy cost beyond that of the task-level private optimizer.

Scope. The result permits adaptive dependence on earlier privatized outputs; it does not require independent models or task mechanisms. Its record-level guarantee relies on each protected record belonging to only one task and on subsequent tasks accessing that record only through privatized outputs. It does not directly provide user-level privacy when one user contributes multiple records across tasks.

C.2 Proof of the Stable–Plastic Geometry

Proof of Theorem 2. Fix a task $t \ge 2$, with nonnegative integer ranks r_S, r_P satisfying $r_S + r_P = r \le n$. We use the bases $\boldsymbol{A}_S^t \in \mathbb{R}^{r_S\times n}$ and $\boldsymbol{A}_P^t \in \mathbb{R}^{r_P\times n}$ constructed in the main text. Define

$$\boldsymbol{P}_S^t = (\boldsymbol{A}_S^t)^\top \boldsymbol{A}_S^t, \qquad \boldsymbol{P}_P^t = (\boldsymbol{A}_P^t)^\top \boldsymbol{A}_P^t, \qquad \boldsymbol{Q}_S^t = \boldsymbol{I}_n - \boldsymbol{P}_S^t.$$

For an update $\boldsymbol{U} \in \mathbb{R}^{m\times n}$, recall that $\mathcal{I}_t(\boldsymbol{U}) = \mathrm{tr}\big(\boldsymbol{U}\boldsymbol{H}^{t-1}\boldsymbol{U}^\top\big)$.

The stable basis consists of orthonormal eigenvectors, so $\boldsymbol{A}_S^t(\boldsymbol{A}_S^t)^\top = \boldsymbol{I}_{r_S}$. Let $\boldsymbol{U}_\perp^t \in \mathbb{R}^{n\times(n-r_S)}$ have orthonormal columns spanning $(\mathcal{S}^t)^\perp$. Then

$$(\boldsymbol{U}_\perp^t)^\top \boldsymbol{U}_\perp^t = \boldsymbol{I}_{n-r_S}, \qquad \boldsymbol{U}_\perp^t(\boldsymbol{U}_\perp^t)^\top = \boldsymbol{Q}_S^t.$$

The plastic construction uses

$$\boldsymbol{C}_P^t = (\boldsymbol{U}_\perp^t)^\top (\boldsymbol{W}^0)^\top \boldsymbol{W}^0 \boldsymbol{U}_\perp^t, \qquad \boldsymbol{A}_P^t = (\boldsymbol{Z}_{P,1:r_P}^t)^\top (\boldsymbol{U}_\perp^t)^\top,$$

where $\boldsymbol{Z}_{P,1:r_P}^t \in \mathbb{R}^{(n-r_S)\times r_P}$ contains orthonormal eigenvectors associated with the r_P largest eigenvalues of $\boldsymbol{C}_P^t$. Therefore, $\boldsymbol{A}_P^t(\boldsymbol{A}_P^t)^\top = \boldsymbol{I}_{r_P}$, $\boldsymbol{A}_P^t(\boldsymbol{A}_S^t)^\top = \boldsymbol{0}$. It follows that

$$\boldsymbol{A}^t(\boldsymbol{A}^t)^\top = \boldsymbol{I}_r, \qquad \boldsymbol{P}^t := (\boldsymbol{A}^t)^\top \boldsymbol{A}^t = \boldsymbol{P}_S^t + \boldsymbol{P}_P^t.$$

In particular, $\boldsymbol{P}_S^t$ and $\boldsymbol{P}_P^t$ are orthogonal projectors with $\boldsymbol{P}_S^t \boldsymbol{P}_P^t = \boldsymbol{0}$.

(i) Historical energy capture. Write the eigendecomposition

$$\boldsymbol{H}^{t-1} = \boldsymbol{V}_H^t\, \mathrm{diag}(h_1^t, \dots, h_n^t)(\boldsymbol{V}_H^t)^\top, \qquad h_1^t \ge \cdots \ge h_n^t \ge 0.$$

Let $\boldsymbol{v}_{H,j}^t$ denote the j-th column of $\boldsymbol{V}_H^t$. Choose this orthonormal eigenbasis consistently with the stable construction, so that $\boldsymbol{A}_S^t = (\boldsymbol{V}_{H,1:r_S}^t)^\top$, including when eigenvalues are repeated.

For any $\boldsymbol{U} \in \mathbb{R}^{r_S\times n}$ satisfying $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}_{r_S}$,

$$\mathrm{tr}\big(\boldsymbol{U}\boldsymbol{H}^{t-1}\boldsymbol{U}^\top\big) = \sum_{j=1}^n h_j^t d_j, \qquad d_j := \|\boldsymbol{U}\boldsymbol{v}_{H,j}^t\|_2^2.$$

Since $\boldsymbol{U}^\top\boldsymbol{U}$ is an orthogonal projector of rank r_S, $0 \le d_j \le 1, \sum_{j=1}^n d_j = \mathrm{tr}(\boldsymbol{U}^\top\boldsymbol{U}) = r_S$. For $1 \le r_S < n$, these relations imply

$$\begin{aligned}\sum_{j=1}^{n} h_j^t d_j &\le \sum_{j=1}^{r_S} h_j^t d_j + h_{r_S}^t \sum_{j=r_S+1}^{n} d_j \\ &= \sum_{j=1}^{r_S} \left[h_j^t d_j + h_{r_S}^t (1-d_j)\right] \\ &\le \sum_{j=1}^{r_S} h_j^t.\end{aligned}$$

The cases $r_S = 0$ and $r_S = n$ follow directly. The selected stable basis attains this upper bound, and hence, for $\boldsymbol{U} \in \mathbb{R}^{r_S \times n}$,

$$\max_{\boldsymbol{U}} \quad \mathrm{tr}\big(\boldsymbol{U}\boldsymbol{H}^{t-1}\boldsymbol{U}^\top\big) = \mathrm{tr}\big(\boldsymbol{A}_S^t\boldsymbol{H}^{t-1}(\boldsymbol{A}_S^t)^\top\big) = \sum_{j=1}^{r_S} h_j^t, \quad \text{s.t.} \quad \boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}_{r_S}.$$

For $r_S < n$, the complementary projector satisfies

$$\boldsymbol{Q}_S^t\boldsymbol{H}^{t-1}\boldsymbol{Q}_S^t = \sum_{j=r_S+1}^{n} h_j^t \boldsymbol{v}_{H,j}^t (\boldsymbol{v}_{H,j}^t)^\top \preceq h_{r_S+1}^t \boldsymbol{Q}_S^t.$$

Because $\boldsymbol{A}_P^t\boldsymbol{Q}_S^t = \boldsymbol{A}_P^t$,

$$\boldsymbol{A}_P^t\boldsymbol{H}^{t-1}(\boldsymbol{A}_P^t)^\top \preceq h_{r_S+1}^t \boldsymbol{I}_{r_P}.$$

Consequently, for every $\boldsymbol{B}_P^t \in \mathbb{R}^{m \times r_P}$,

$$\begin{aligned}\mathcal{I}_t(\boldsymbol{B}_P^t\boldsymbol{A}_P^t) &= \mathrm{tr}\big(\boldsymbol{B}_P^t\boldsymbol{A}_P^t\boldsymbol{H}^{t-1}(\boldsymbol{A}_P^t)^\top(\boldsymbol{B}_P^t)^\top\big) \\ &\le h_{r_S+1}^t \,\mathrm{tr}\big(\boldsymbol{B}_P^t(\boldsymbol{B}_P^t)^\top\big) \\ &= h_{r_S+1}^t \|\boldsymbol{B}_P^t\|_F^2.\end{aligned}$$

If $r_S = n$, then $r_P = 0$ and the plastic update is zero; the overlap bound is interpreted as zero without introducing h_{n+1}^t.

(ii) Complementary pretrained response. Let $\mu_1^t \ge \cdots \ge \mu_{n-r_S}^t \ge 0$ be the eigenvalues of $\boldsymbol{C}_P^t$. Consider any $\boldsymbol{U} \in \mathbb{R}^{r_P \times n}$ satisfying

$$\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}_{r_P}, \qquad \boldsymbol{U}(\boldsymbol{A}_S^t)^\top = \boldsymbol{0}.$$

Set

$$\widetilde{\boldsymbol{U}} := \boldsymbol{U}\boldsymbol{U}_\perp^t \in \mathbb{R}^{r_P \times (n-r_S)}.$$

The orthogonality constraint gives $\boldsymbol{U}\boldsymbol{Q}_S^t = \boldsymbol{U}$, and therefore

$$\boldsymbol{U} = \widetilde{\boldsymbol{U}}(\boldsymbol{U}_\perp^t)^\top, \qquad \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{U}}^\top = \boldsymbol{U}\boldsymbol{Q}_S^t\boldsymbol{U}^\top = \boldsymbol{I}_{r_P}.$$

Conversely, every row-orthonormal $\widetilde{\boldsymbol{U}}$ of this dimension produces an admissible $\boldsymbol{U}$ through this expression. Thus,

$$\begin{aligned}\|\boldsymbol{W}^0\boldsymbol{U}^\top\|_F^2 &= \mathrm{tr}\big(\boldsymbol{U}(\boldsymbol{W}^0)^\top\boldsymbol{W}^0\boldsymbol{U}^\top\big) \\ &= \mathrm{tr}\Big(\widetilde{\boldsymbol{U}}\boldsymbol{C}_P^t\widetilde{\boldsymbol{U}}^\top\Big) \\ &\le \sum_{j=1}^{r_P} \mu_j^t,\end{aligned}$$

where the last step applies the eigenvalue argument from part (i) to $\boldsymbol{C}_P^t$. Equality holds for $\widetilde{\boldsymbol{U}} = (\boldsymbol{Z}_{P,1:r_P}^t)^\top$, which yields the selected plastic basis. Hence for $\boldsymbol{U} \in \mathbb{R}^{r_P \times n}$,

$$\max_{\boldsymbol{U}} \|\boldsymbol{W}^0\boldsymbol{U}^\top\|_F^2 = \|\boldsymbol{W}^0(\boldsymbol{A}_P^t)^\top\|_F^2 = \sum_{j=1}^{r_P} \mu_j^t, \quad \text{s.t. } \boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}_{r_P},\ \boldsymbol{U}(\boldsymbol{A}_S^t)^\top = \boldsymbol{0}.$$

Repeated or zero eigenvalues do not change this optimal value: an orthonormal selection of the corresponding eigenvectors in the restricted space still attains it.

(iii) Additional adaptation capacity. For every $\boldsymbol{Y} \in \mathbb{R}^{m\times r}$, row orthonormality gives

$$\|\boldsymbol{Y}\boldsymbol{A}^t\|_F^2 = \mathrm{tr}\big(\boldsymbol{Y}\boldsymbol{A}^t(\boldsymbol{A}^t)^\top\boldsymbol{Y}^\top\big) = \|\boldsymbol{Y}\|_F^2.$$

This proves the claimed norm identity.

For a target update $\boldsymbol{U} \in \mathbb{R}^{m\times n}$ and coefficients $\boldsymbol{X} \in \mathbb{R}^{m\times r}$, decompose

$$\boldsymbol{U} - \boldsymbol{X}\boldsymbol{A}^t = \boldsymbol{U}(\boldsymbol{I}_n - \boldsymbol{P}^t) + \big(\boldsymbol{U}(\boldsymbol{A}^t)^\top - \boldsymbol{X}\big)\,\boldsymbol{A}^t.$$

The two terms are orthogonal in the Frobenius inner product because $(\boldsymbol{I}_n - \boldsymbol{P}^t)(\boldsymbol{A}^t)^\top = \boldsymbol{0}$. Applying the norm identity gives

$$\|\boldsymbol{U} - \boldsymbol{X}\boldsymbol{A}^t\|_F^2 = \|\boldsymbol{U}(\boldsymbol{I}_n - \boldsymbol{P}^t)\|_F^2 + \|\boldsymbol{U}(\boldsymbol{A}^t)^\top - \boldsymbol{X}\|_F^2.$$

The unique minimizing coefficient is therefore $\boldsymbol{X} = \boldsymbol{U}(\boldsymbol{A}^t)^\top$, with

$$\min_{\boldsymbol{X}\in\mathbb{R}^{m\times r}} \|\boldsymbol{U} - \boldsymbol{X}\boldsymbol{A}^t\|_F^2 = \|\boldsymbol{U}(\boldsymbol{I}_n - \boldsymbol{P}^t)\|_F^2.$$

The same argument applied to the stable basis yields

$$\min_{\boldsymbol{X}_S\in\mathbb{R}^{m\times r_S}} \|\boldsymbol{U} - \boldsymbol{X}_S\boldsymbol{A}_S^t\|_F^2 = \|\boldsymbol{U}(\boldsymbol{I}_n - \boldsymbol{P}_S^t)\|_F^2.$$

Since $\boldsymbol{I}_n - \boldsymbol{P}_S^t = (\boldsymbol{I}_n - \boldsymbol{P}^t) + \boldsymbol{P}_P^t$ is a sum of mutually orthogonal projectors,

$$\|\boldsymbol{U}(\boldsymbol{I}_n - \boldsymbol{P}_S^t)\|_F^2 = \|\boldsymbol{U}(\boldsymbol{I}_n - \boldsymbol{P}^t)\|_F^2 + \|\boldsymbol{U}\boldsymbol{P}_P^t\|_F^2.$$

Moreover,

$$\|\boldsymbol{U}\boldsymbol{P}_P^t\|_F^2 = \|(\boldsymbol{U}(\boldsymbol{A}_P^t)^\top)\boldsymbol{A}_P^t\|_F^2 = \|\boldsymbol{U}(\boldsymbol{A}_P^t)^\top\|_F^2,$$

using row orthonormality of $\boldsymbol{A}_P^t$. Subtracting the optimal residuals gives

$$\min_{\boldsymbol{X}_S\in\mathbb{R}^{m\times r_S}} \|\boldsymbol{U} - \boldsymbol{X}_S\boldsymbol{A}_S^t\|_F^2 - \min_{\boldsymbol{X}\in\mathbb{R}^{m\times r}} \|\boldsymbol{U} - \boldsymbol{X}\boldsymbol{A}^t\|_F^2 = \|\boldsymbol{U}(\boldsymbol{A}_P^t)^\top\|_F^2.$$

This establishes all claims of the theorem. Empty branches are interpreted using empty matrices and zero sums. □

Interpretation. The approximation identity compares the same stable basis with and without the selected plastic branch. It quantifies the additional representational capacity provided by that branch, rather than superiority over all alternative bases of total rank r.

C.3 Proof of the Recalibration Guarantee

Proof of Theorem 3. Fix a task $t \geq 2$ and condition on the realized private coefficient matrix $\boldsymbol{B}^t = [\boldsymbol{B}_S^t, \boldsymbol{B}_P^t]$ and history $\boldsymbol{H}^{t-1}$. For each stable direction j, let $\boldsymbol{b}_{S,j}^t \in \mathbb{R}^m$ denote the j-th column of $\boldsymbol{B}_S^t$, and recall $c_j^t = \|\boldsymbol{b}_{S,j}^t\|_2^2, \lambda_j^t = \frac{c_j^t}{h_j^t + c_j^t + \tau}, \boldsymbol{\Lambda}_S^t = \mathrm{diag}(\lambda_1^t, \ldots, \lambda_{r_S}^t)$. Here $h_j^t \geq 0$ is the historical eigenvalue associated with the j-th stable basis vector, and $\tau > 0$. Thus $0 \leq \lambda_j^t < 1$.

Define the recalibrated coefficient matrix $\widehat{\boldsymbol{B}}^t = [\boldsymbol{B}_S^t\boldsymbol{\Lambda}_S^t, \boldsymbol{B}_P^t]$. The updates before and after recalibration are $\Delta\boldsymbol{W}_{\mathrm{priv}}^t = \boldsymbol{B}^t\boldsymbol{A}^t$ and $\Delta\boldsymbol{W}^t = \widehat{\boldsymbol{B}}^t\boldsymbol{A}^t$, respectively.

(i) Unique minimizer of the recalibration objective. Write $\boldsymbol{X} = [\boldsymbol{X}_S, \boldsymbol{X}_P]$, where $\boldsymbol{X}_S \in \mathbb{R}^{m\times r_S}$ and $\boldsymbol{X}_P \in \mathbb{R}^{m\times r_P}$, and let $\boldsymbol{x}_j$ denote the j-th column of $\boldsymbol{X}_S$. The objective in the theorem is

$$J_t(\boldsymbol{X}) = \frac{1}{2}\|\boldsymbol{X} - \boldsymbol{B}^t\|_F^2 + \frac{1}{2}\sum_{j:c_j^t>0} \frac{h_j^t + \tau}{c_j^t}\|\boldsymbol{x}_j\|_2^2,$$

subject to $\boldsymbol{x}_j = \boldsymbol{0}$ whenever $c_j^t = 0$. Because the Frobenius norm separates over columns, each free stable column minimizes

$$J_{t,j}(\boldsymbol{x}_j) = \frac{1}{2}\|\boldsymbol{x}_j - \boldsymbol{b}_{S,j}^t\|_2^2 + \frac{h_j^t + \tau}{2c_j^t}\|\boldsymbol{x}_j\|_2^2.$$

For $c_j^t > 0$, its gradient and Hessian are

$$\nabla J_{t,j}(\boldsymbol{x}_j) = \left(1 + \frac{h_j^t + \tau}{c_j^t}\right)\boldsymbol{x}_j - \boldsymbol{b}_{S,j}^t, \qquad \nabla^2 J_{t,j} = \left(1 + \frac{h_j^t + \tau}{c_j^t}\right)\boldsymbol{I}_m \succ 0.$$

Setting the gradient to zero therefore gives the unique minimizer

$$\boldsymbol{x}_j = \frac{c_j^t}{c_j^t + h_j^t + \tau}\boldsymbol{b}_{S,j}^t = \lambda_j^t \boldsymbol{b}_{S,j}^t.$$

If $c_j^t = 0$, then $\boldsymbol{b}_{S,j}^t = \boldsymbol{0}$ and $\lambda_j^t = 0$, so the imposed constraint $\boldsymbol{x}_j = \boldsymbol{0}$ agrees with the same rule.

The plastic block contributes only $\frac{1}{2}\|\boldsymbol{X}_P - \boldsymbol{B}_P^t\|_F^2$, whose unique minimizer is $\boldsymbol{X}_P = \boldsymbol{B}_P^t$. Combining these columnwise solutions proves that $\widehat{\boldsymbol{B}}^t$ is the unique minimizer of the full constrained objective.

(ii) Plastic coefficients and update norms. The orthogonality relations established in Theorem 2 give $\boldsymbol{A}_S^t(\boldsymbol{A}_P^t)^\top = \boldsymbol{0}, \qquad \boldsymbol{A}_P^t(\boldsymbol{A}_P^t)^\top = \boldsymbol{I}_{r_P}$. Consequently,

$$\begin{aligned}\Delta \boldsymbol{W}^t(\boldsymbol{A}_P^t)^\top &= \boldsymbol{B}_S^t \boldsymbol{\Lambda}_S^t \boldsymbol{A}_S^t(\boldsymbol{A}_P^t)^\top + \boldsymbol{B}_P^t \boldsymbol{A}_P^t(\boldsymbol{A}_P^t)^\top \\ &= \boldsymbol{B}_P^t.\end{aligned}$$

Thus recalibration preserves the plastic coefficients exactly.

Since $\boldsymbol{A}^t(\boldsymbol{A}^t)^\top = \boldsymbol{I}_r$, multiplication by $\boldsymbol{A}^t$ preserves the Frobenius norm of coefficient matrices. Therefore,

$$\begin{aligned}\|\Delta \boldsymbol{W}_{\text{priv}}^t - \Delta \boldsymbol{W}^t\|_F^2 &= \|\boldsymbol{B}^t - \widehat{\boldsymbol{B}}^t\|_F^2 \\ &= \sum_{j=1}^{r_S} c_j^t (1 - \lambda_j^t)^2. \end{aligned} \tag{10}$$

Likewise,

$$\|\Delta \boldsymbol{W}_{\text{priv}}^t\|_F^2 = \sum_{j=1}^{r_S} c_j^t + \|\boldsymbol{B}_P^t\|_F^2, \qquad \|\Delta \boldsymbol{W}^t\|_F^2 = \sum_{j=1}^{r_S} c_j^t (\lambda_j^t)^2 + \|\boldsymbol{B}_P^t\|_F^2.$$

Subtracting yields

$$\|\Delta \boldsymbol{W}_{\text{priv}}^t\|_F^2 - \|\Delta \boldsymbol{W}^t\|_F^2 = \sum_{j=1}^{r_S} c_j^t \big(1 - (\lambda_j^t)^2\big) \geq 0. \tag{11}$$

Hence the update norm cannot increase.

(iii) Historical overlap reduction. Define $\boldsymbol{D}_H^t = \operatorname{diag}(h_1^t, \ldots, h_{r_S}^t)$. Because the stable rows are eigenvectors of $\boldsymbol{H}^{t-1}$,

$$\boldsymbol{A}_S^t \boldsymbol{H}^{t-1} = \boldsymbol{D}_H^t \boldsymbol{A}_S^t.$$

Combining this identity with row orthonormality and stable–plastic orthogonality gives

$$\boldsymbol{A}_S^t \boldsymbol{H}^{t-1}(\boldsymbol{A}_S^t)^\top = \boldsymbol{D}_H^t, \qquad \boldsymbol{A}_S^t \boldsymbol{H}^{t-1}(\boldsymbol{A}_P^t)^\top = \boldsymbol{0}.$$

The transposed cross term also vanishes because $\boldsymbol{H}^{t-1}$ is symmetric.

Recall $\mathcal{I}_t(\boldsymbol{U}) = \operatorname{tr}(\boldsymbol{U}\boldsymbol{H}^{t-1}\boldsymbol{U}^\top)$. Expanding this expression for the two updates, the vanishing cross terms imply

$$\begin{aligned}\mathcal{I}_t(\Delta \boldsymbol{W}_{\text{priv}}^t) &= \operatorname{tr}\big(\boldsymbol{B}_S^t \boldsymbol{D}_H^t (\boldsymbol{B}_S^t)^\top\big) + \mathcal{I}_t(\boldsymbol{B}_P^t \boldsymbol{A}_P^t) \\ &= \sum_{j=1}^{r_S} h_j^t c_j^t + \mathcal{I}_t(\boldsymbol{B}_P^t \boldsymbol{A}_P^t), \\ \mathcal{I}_t(\Delta \boldsymbol{W}^t) &= \operatorname{tr}\big(\boldsymbol{B}_S^t \boldsymbol{\Lambda}_S^t \boldsymbol{D}_H^t \boldsymbol{\Lambda}_S^t (\boldsymbol{B}_S^t)^\top\big) + \mathcal{I}_t(\boldsymbol{B}_P^t \boldsymbol{A}_P^t) \\ &= \sum_{j=1}^{r_S} h_j^t c_j^t (\lambda_j^t)^2 + \mathcal{I}_t(\boldsymbol{B}_P^t \boldsymbol{A}_P^t).\end{aligned}$$

The plastic contribution is identical in both expressions and cancels upon subtraction:

$$\mathcal{I}_t(\Delta \boldsymbol{W}^t_{\text{priv}}) - \mathcal{I}_t(\Delta \boldsymbol{W}^t) = \sum_{j=1}^{r_S} h_j^t c_j^t \big(1 - (\lambda_j^t)^2\big) \geq 0. \tag{12}$$

Every summand is nonnegative since $h_j^t, c_j^t \geq 0$ and $0 \leq \lambda_j^t < 1$. The decrease is strict whenever $h_j^t c_j^t > 0$ for at least one stable direction. This proves all claims of the theorem. Empty branches are interpreted using empty matrices and zero sums. □

Interpretation. These identities hold for each realized private update, without assuming that c_j^t is a noise-free measure of task relevance. They establish exact preservation of plastic coefficients and non-increasing historical overlap; their implications for old-task loss are addressed separately in Theorem 4.

C.4 Proof of the Conditional Old-Task Loss Bound

Proof of Theorem 4. Fix a task $t \geq 2$, the pre-merge weight $\boldsymbol{W}^{t-1}$, and the realized history and private coefficients. Consider the two updates $\Delta \boldsymbol{W}^t_{\text{priv}}$ and $\Delta \boldsymbol{W}^t$ before and after recalibration. Throughout the proof, both updates are evaluated from the same pre-merge weight.

For matrices of the same dimension, write $\langle \boldsymbol{X}, \boldsymbol{Y} \rangle_F = \operatorname{tr}(\boldsymbol{X}^\top \boldsymbol{Y})$ for the Frobenius inner product. The notation $\nabla^2 F_{\text{old}}(\boldsymbol{W})[\boldsymbol{V}, \boldsymbol{V}]$ denotes the Hessian quadratic form in direction $\boldsymbol{V}$. Recall

$$\mathcal{I}_t(\boldsymbol{V}) = \operatorname{tr}\big(\boldsymbol{V} \boldsymbol{H}^{t-1} \boldsymbol{V}^\top\big).$$

By the theorem's assumptions,

$$\|\nabla F_{\text{old}}(\boldsymbol{W}^{t-1})\|_F \leq \nu.$$

Moreover, the same constants $\kappa, \rho \geq 0$ satisfy

$$\nabla^2 F_{\text{old}}(\boldsymbol{W}^{t-1} + s\boldsymbol{U})[\boldsymbol{V}, \boldsymbol{V}] \leq \kappa \mathcal{I}_t(\boldsymbol{V}) + \rho \|\boldsymbol{V}\|_F^2$$

for both $\boldsymbol{U} \in \{\Delta \boldsymbol{W}^t_{\text{priv}}, \Delta \boldsymbol{W}^t\}$, every $s \in [0, 1]$, and every $\boldsymbol{V} \in \mathbb{R}^{m \times n}$. The common constants and history define the same upper-bound function for both updates.

(i) Upper bound on the old-task loss increase. For either admissible update $\boldsymbol{U}$, define the scalar function

$$f_{\boldsymbol{U}}(s) = F_{\text{old}}(\boldsymbol{W}^{t-1} + s\boldsymbol{U}), \qquad s \in [0, 1].$$

The differentiability assumption gives

$$f'_{\boldsymbol{U}}(0) = \langle \nabla F_{\text{old}}(\boldsymbol{W}^{t-1}), \boldsymbol{U} \rangle_F, \qquad f''_{\boldsymbol{U}}(s) = \nabla^2 F_{\text{old}}(\boldsymbol{W}^{t-1} + s\boldsymbol{U})[\boldsymbol{U}, \boldsymbol{U}].$$

Taylor's formula with integral remainder yields

$$\begin{aligned} &F_{\text{old}}(\boldsymbol{W}^{t-1} + \boldsymbol{U}) - F_{\text{old}}(\boldsymbol{W}^{t-1}) \\ &\quad = \langle \nabla F_{\text{old}}(\boldsymbol{W}^{t-1}), \boldsymbol{U} \rangle_F \\ &\qquad + \int_0^1 (1 - s) \nabla^2 F_{\text{old}}(\boldsymbol{W}^{t-1} + s\boldsymbol{U})[\boldsymbol{U}, \boldsymbol{U}]\, ds. \end{aligned}$$

By Cauchy–Schwarz and the gradient bound,

$$\langle \nabla F_{\text{old}}(\boldsymbol{W}^{t-1}), \boldsymbol{U} \rangle_F \leq \|\nabla F_{\text{old}}(\boldsymbol{W}^{t-1})\|_F \|\boldsymbol{U}\|_F \leq \nu \|\boldsymbol{U}\|_F.$$

Applying the curvature bound with $\boldsymbol{V} = \boldsymbol{U}$ and using $\int_0^1 (1 - s)\, ds = 1/2$, we obtain

$$\begin{aligned} &\int_0^1 (1 - s) \nabla^2 F_{\text{old}}(\boldsymbol{W}^{t-1} + s\boldsymbol{U})[\boldsymbol{U}, \boldsymbol{U}]\, ds \\ &\qquad \leq \frac{\kappa}{2} \mathcal{I}_t(\boldsymbol{U}) + \frac{\rho}{2} \|\boldsymbol{U}\|_F^2. \end{aligned}$$

Combining these inequalities proves

$$F_{\text{old}}(\boldsymbol{W}^{t-1} + \boldsymbol{U}) - F_{\text{old}}(\boldsymbol{W}^{t-1}) \leq \mathcal{U}_t(\boldsymbol{U}),$$

where

$$\mathcal{U}_t(\boldsymbol{U}) = \nu \|\boldsymbol{U}\|_F + \frac{\kappa}{2} \mathcal{I}_t(\boldsymbol{U}) + \frac{\rho}{2} \|\boldsymbol{U}\|_F^2.$$

This inequality holds for both updates.

(ii) Comparison before and after recalibration. Equations 11 and 12 give

$$\|\Delta \boldsymbol{W}^t_{\text{priv}}\|_F^2 - \|\Delta \boldsymbol{W}^t\|_F^2 = \sum_{j=1}^{r_S} c_j^t\big(1-(\lambda_j^t)^2\big) \geq 0,$$

$$\mathcal{I}_t(\Delta \boldsymbol{W}^t_{\text{priv}}) - \mathcal{I}_t(\Delta \boldsymbol{W}^t) = \sum_{j=1}^{r_S} h_j^t c_j^t\big(1-(\lambda_j^t)^2\big) \geq 0.$$

Since norms are nonnegative, the first inequality also implies

$$\|\Delta \boldsymbol{W}^t\|_F \leq \|\Delta \boldsymbol{W}^t_{\text{priv}}\|_F.$$

Subtracting the two upper bounds and substituting the identities above yields

$$\begin{aligned}\mathcal{U}_t(\Delta \boldsymbol{W}^t_{\text{priv}}) &- \mathcal{U}_t(\Delta \boldsymbol{W}^t) \\ &= \nu\left(\|\Delta \boldsymbol{W}^t_{\text{priv}}\|_F - \|\Delta \boldsymbol{W}^t\|_F\right) \\ &\quad + \frac{1}{2}\sum_{j=1}^{r_S}(\kappa h_j^t + \rho) c_j^t\big(1-(\lambda_j^t)^2\big) \geq 0. \end{aligned} \tag{13}$$

Every term is nonnegative because $\nu, \kappa, \rho, h_j^t, c_j^t \geq 0$ and $0 \leq \lambda_j^t < 1$. Therefore,

$$\mathcal{U}_t(\Delta \boldsymbol{W}^t) \leq \mathcal{U}_t(\Delta \boldsymbol{W}^t_{\text{priv}}),$$

which proves the theorem. □

Interpretation. The theorem compares a common upper bound for two updates at the same pre-merge model. It does not imply an ordering of their actual old-task losses or of complete training trajectories with different histories. The connection between historical overlap and old-task loss enters through the stated curvature assumption.

Extension to multiple adapted layers. Let $\ell = 1, \ldots, L_{\text{ad}}$ index the adapted weight matrices, and collect their increments as $\boldsymbol{U} = (\boldsymbol{U}_\ell)_{\ell=1}^{L_{\text{ad}}}$. On this product space, define

$$\|\boldsymbol{U}\|_{\text{joint}}^2 = \sum_{\ell=1}^{L_{\text{ad}}} \|\boldsymbol{U}_\ell\|_F^2, \qquad \mathcal{I}_t^{\text{joint}}(\boldsymbol{U}) = \sum_{\ell=1}^{L_{\text{ad}}} \operatorname{tr}\big(\boldsymbol{U}_\ell \boldsymbol{H}_\ell^{t-1} \boldsymbol{U}_\ell^\top\big).$$

Here $\boldsymbol{H}_\ell^{t-1}$ is the history matrix for layer ℓ. Let $\boldsymbol{W}^{t-1}$ now denote the collection of pre-merge matrices, with all other model parameters held fixed in the comparison.

Suppose the old-task loss is twice continuously differentiable along both joint update segments, its joint gradient satisfies

$$\|\nabla F_{\text{old}}(\boldsymbol{W}^{t-1})\|_{\text{joint}} \leq \nu,$$

and its full Hessian satisfies

$$\nabla^2 F_{\text{old}}(\boldsymbol{W}^{t-1} + s\boldsymbol{U})[\boldsymbol{V}, \boldsymbol{V}] \leq \kappa \mathcal{I}_t^{\text{joint}}(\boldsymbol{V}) + \rho \|\boldsymbol{V}\|_{\text{joint}}^2$$

for both joint updates, all $s \in [0,1]$, and every joint perturbation $\boldsymbol{V}$. This is a bound on the full Hessian, including cross-layer terms, rather than separate bounds on individual layers.

The same Taylor argument gives the joint upper bound

$$\mathcal{U}_t^{\text{joint}}(\boldsymbol{U}) = \nu\|\boldsymbol{U}\|_{\text{joint}} + \frac{\kappa}{2}\mathcal{I}_t^{\text{joint}}(\boldsymbol{U}) + \frac{\rho}{2}\|\boldsymbol{U}\|_{\text{joint}}^2.$$

Applying the recalibration identities to each layer and summing shows that neither the joint squared norm nor the joint historical overlap increases. Hence

$$\mathcal{U}_t^{\text{joint}}(\Delta \boldsymbol{W}^t) \leq \mathcal{U}_t^{\text{joint}}(\Delta \boldsymbol{W}^t_{\text{priv}}).$$

Thus the single-matrix argument extends to multiple adapted layers under the stated joint assumptions.

D SUPPLEMENTARY ANALYSIS OF THE BILINEAR NOISE INTERACTION

This section supplements Observation 2 and Section 4.2 by quantifying the fresh-noise energy in one DP-SGD step.

Setup. Condition on the optimization state and current minibatch before drawing fresh noise. Let $\bar{\boldsymbol{B}} \in \mathbb{R}^{m\times r}$ and $\bar{\boldsymbol{A}} \in \mathbb{R}^{r\times n}$ be the factors after the clipped-gradient update without this step's noise. Let $\boldsymbol{N}_B \in \mathbb{R}^{m\times r}$ and $\boldsymbol{N}_A \in \mathbb{R}^{r\times n}$ have mutually independent centered Gaussian entries with conditional variances v_B^2 and v_A^2, respectively. All expectations below are over these fresh perturbations, conditional on the fixed state and minibatch.

Proposition 1 (Single-step noise energy) *Under the preceding assumptions,*

$$\begin{aligned}\mathbb{E}\left\|(\bar{\boldsymbol{B}}+\boldsymbol{N}_B)(\bar{\boldsymbol{A}}+\boldsymbol{N}_A)-\bar{\boldsymbol{B}}\bar{\boldsymbol{A}}\right\|_F^2 \\ = nv_A^2\|\bar{\boldsymbol{B}}\|_F^2 + mv_B^2\|\bar{\boldsymbol{A}}\|_F^2 + mnr\,v_A^2v_B^2.\end{aligned} \tag{14}$$

For a fixed basis $\boldsymbol{A}^t$ satisfying $\boldsymbol{A}^t(\boldsymbol{A}^t)^\top = \boldsymbol{I}_r$,

$$\|\boldsymbol{N}_B\boldsymbol{A}^t\|_F^2 = \|\boldsymbol{N}_B\|_F^2, \qquad \mathbb{E}\|\boldsymbol{N}_B\boldsymbol{A}^t\|_F^2 = mr\,v_B^2. \tag{15}$$

Proof of Proposition 1. The joint perturbation expands as $\bar{\boldsymbol{B}}\boldsymbol{N}_A + \boldsymbol{N}_B\bar{\boldsymbol{A}} + \boldsymbol{N}_B\boldsymbol{N}_A$. All pairwise Frobenius inner products have zero expectation: the first two terms are independently centered; for the first and third, condition additionally on $\boldsymbol{N}_A$ and average over $\boldsymbol{N}_B$; for the second and third, reverse this conditioning. Thus the expected squared norm is the sum of the three expected squared norms.

Using $\mathbb{E}[\boldsymbol{N}_A\boldsymbol{N}_A^\top] = nv_A^2\boldsymbol{I}_r$ and $\mathbb{E}[\boldsymbol{N}_B^\top\boldsymbol{N}_B] = mv_B^2\boldsymbol{I}_r$ gives

$$\begin{aligned}\mathbb{E}\|\bar{\boldsymbol{B}}\boldsymbol{N}_A\|_F^2 &= nv_A^2\|\bar{\boldsymbol{B}}\|_F^2,\\ \mathbb{E}\|\boldsymbol{N}_B\bar{\boldsymbol{A}}\|_F^2 &= mv_B^2\|\bar{\boldsymbol{A}}\|_F^2,\\ \mathbb{E}\|\boldsymbol{N}_B\boldsymbol{N}_A\|_F^2 &= nv_A^2\mathbb{E}\|\boldsymbol{N}_B\|_F^2 = mnr\,v_A^2v_B^2.\end{aligned}$$

The last equality follows by first conditioning on $\boldsymbol{N}_B$. Summing proves Eq. 14. For the fixed-basis case, $\|\boldsymbol{N}_B\boldsymbol{A}^t\|_F^2 = \mathrm{tr}(\boldsymbol{N}_B\boldsymbol{A}^t(\boldsymbol{A}^t)^\top\boldsymbol{N}_B^\top) = \|\boldsymbol{N}_B\|_F^2$. Taking expectations proves Eq. 15. □

Interpretation and scope. The term $mnr\,v_A^2v_B^2$ is the expected energy of the bilinear interaction in **Observation 2**. Fixing $\boldsymbol{A}^t$ removes both perturbation terms involving $\boldsymbol{N}_A$, while its row orthonormality preserves the remaining coefficient-noise energy. At $\bar{\boldsymbol{A}} = \boldsymbol{A}^t$ and the same v_B, the joint-factor expected energy exceeds the single-factor value by the nonnegative quantity $nv_A^2\|\bar{\boldsymbol{B}}\|_F^2 + mnr\,v_A^2v_B^2$. The empirical measurements in Observation 2 assess the interaction's relative magnitude during training.

For SGD with learning rate η, clipping norm C, noise multiplier σ, and normalization $b > 0$ fixed under the above conditioning, the parameter-noise standard deviation is $v = \eta\sigma C/b$. Adaptive optimizers require a separate analysis of how gradient noise is transformed. These results concern conditional single-step perturbations, not final model errors or accuracies: clipped gradients and subsequent trajectories may differ between joint-factor and single-factor training. They also do not characterize noise after nonlinear, data-dependent recalibration.

E SUPPLEMENTARY EXPERIMENTAL INFORMATION

E.1 EXPERIMENTAL DETAILS

Data splits and task construction. We follow the training/test splits and class-incremental task construction of He et al. (2026) for the five benchmarks used in our experiments. We evaluate ImageNet-R with $T \in \{5, 10, 20\}$, ImageNet-A with $T \in \{10, 20\}$, and CIFAR-100, CUB, and DomainNet with $T = 10$. Task class sets are disjoint. After each task, we evaluate the model over all classes observed so far, without access to task identity.

Privacy budget. Each disjoint task partition is assigned the same privacy budget (ϵ, δ). Provided that all private-data-dependent operations within each task are covered by this budget and subsequent tasks access earlier partitions only through DP outputs, parallel composition yields the same (ϵ, δ) guarantee for the entire continual learning process.

Baseline selection. We use AB-LoRA and FA-LoRA as controlled baselines, both with rank $r = 32$, matching the total rank of DP-SPA. AB-LoRA trains both low-rank factors, whereas FA-LoRA fixes the input-side factor and trains only the output-side factor. This comparison examines whether the benefits of DP-SPA extend beyond fixing one factor under differential privacy.

Our comparison focuses on LoRA-based adaptation under DP rather than a comprehensive evaluation of continual learning methods. Non-private continual learning results are not directly comparable at a fixed privacy budget. Adapting methods with additional private-data-dependent operations requires accounting for those operations as well as gradient updates, and may require algorithmic changes. We therefore restrict our controlled comparison to these two baselines.

E.2 Additional Results

Comparison of recalibration strategies. To assess whether direction-wise recalibration provides benefits beyond uniform shrinkage, we compare four strategies: (i) no recalibration; (ii) constant scaling of the stable branch by $\gamma = 0.5$; (iii) norm-matched uniform scaling; and (iv) the proposed history-aware recalibration. All variants use the same stable–plastic basis construction and private optimization protocol, with the plastic branch retained without rescaling.

For norm-matched scaling, each LoRA module computes λ_j^t from its own history and private update, but applies a uniform coefficient

$$\gamma^t = \sqrt{\frac{\sum_{j=1}^{r_S} c_j^t (\lambda_j^t)^2}{\sum_{j=1}^{r_S} c_j^t}}, \tag{16}$$

with $\gamma^t = 0$ when the denominator is zero. The row orthonormality of $\boldsymbol{A}_S^t$ ensures $\|\gamma^t \boldsymbol{B}_S^t \boldsymbol{A}_S^t\|_F = \|\boldsymbol{B}_S^t \boldsymbol{\Lambda}_S^t \boldsymbol{A}_S^t\|_F$. This matches the stable-branch update norm at the same pre-merge state while removing direction-wise scaling. Each variant runs the full continual sequence, maintains its own history using its merged updates, and omits recalibration on the first task.

Table 7 reports the results on ImageNet-R and ImageNet-A with $T = 20$. DP-SPA achieves the highest mean last and average accuracies on both datasets. Compared with no recalibration, it improves $\mathcal{A}_{Last}$ by 0.89 and 0.88 percentage points, respectively. Constant scaling underperforms no recalibration on both datasets, indicating that a fixed shrinkage factor does not reliably improve performance. Norm-matched scaling yields modest gains in $\mathcal{A}_{Last}$ but does not consistently improve $\mathcal{A}_{Avg}$. DP-SPA further exceeds norm-matched scaling by 0.57 and 0.64 percentage points in $\mathcal{A}_{Last}$, and by 0.69 and 1.57 points in $\mathcal{A}_{Avg}$, respectively.

Table 7: Comparison of recalibration strategies on ImageNet-R and ImageNet-A with $T = 20$. Constant scaling uses $\gamma = 0.5$, while norm-matched scaling matches the stable-branch update norm of direction-wise recalibration at the same pre-merge state.

Strategy	**ImageNet-R** (T=20)		**ImageNet-A** (T=20)	
	$\mathcal{A}_{Last} \uparrow$	$\mathcal{A}_{Avg} \uparrow$	$\mathcal{A}_{Last} \uparrow$	$\mathcal{A}_{Avg} \uparrow$
No recalibration	$67.25_{\pm 0.69}$	$77.49_{\pm 0.72}$	$47.56_{\pm 0.63}$	$59.68_{\pm 0.55}$
Constant scaling	$65.30_{\pm 0.54}$	$76.83_{\pm 0.61}$	$44.97_{\pm 0.67}$	$58.18_{\pm 0.52}$
Norm-matched scaling	$67.57_{\pm 0.59}$	$77.65_{\pm 0.42}$	$47.80_{\pm 0.49}$	$59.32_{\pm 0.60}$
DP-SPA	$68.14_{\pm 0.75}$	$78.34_{\pm 0.67}$	$48.44_{\pm 0.84}$	$60.89_{\pm 0.78}$

Sensitivity to stable–plastic rank allocation. Table 8 compares different allocations with fixed total rank $r = 32$, $T = 20$, and $\epsilon = 3$. The best mean accuracies are obtained with $(r_S, r_P) = (16, 16)$ on ImageNet-R and $(8, 24)$ on ImageNet-A, indicating dataset-dependent allocation preferences. All evaluated mixed allocations outperform both single-branch configurations on both met-

rics, supporting the complementary roles of stable and plastic directions. We use $(16, 16)$ as a common default across datasets.

Table 8: Sensitivity to stable–plastic rank allocation on ImageNet-R and ImageNet-A with $T = 20$, $\epsilon = 3$, and fixed total rank $r_S + r_P = 32$.

(r_S, r_P)	**ImageNet-R** (T=20)		**ImageNet-A** (T=20)	
	$\mathcal{A}_{Last} \uparrow$	$\mathcal{A}_{Avg} \uparrow$	$\mathcal{A}_{Last} \uparrow$	$\mathcal{A}_{Avg} \uparrow$
$(0, 32)$	$65.80_{\pm 0.74}$	$76.32_{\pm 0.81}$	$45.19_{\pm 0.90}$	$58.24_{\pm 0.68}$
$(8, 24)$	$67.91_{\pm 0.82}$	$77.63_{\pm 0.78}$	$49.12_{\pm 0.95}$	$61.06_{\pm 0.77}$
$(16, 16)$	$68.14_{\pm 0.75}$	$78.34_{\pm 0.67}$	$48.44_{\pm 0.84}$	$60.89_{\pm 0.78}$
$(24, 8)$	$67.93_{\pm 0.30}$	$77.83_{\pm 0.28}$	$47.56_{\pm 0.46}$	$59.97_{\pm 0.34}$
$(32, 0)$	$67.76_{\pm 0.60}$	$77.15_{\pm 0.46}$	$47.05_{\pm 0.65}$	$59.43_{\pm 0.52}$

Performance on ImageNet-A. Figure 6 compares overall, old-task, and new-task accuracy across 20 incremental tasks on ImageNet-A. DP-SPA achieves higher overall and old-task accuracy through most stages while maintaining new-task accuracy comparable to FA-LoRA, suggesting that improved retention is the main contributor to its overall advantage.

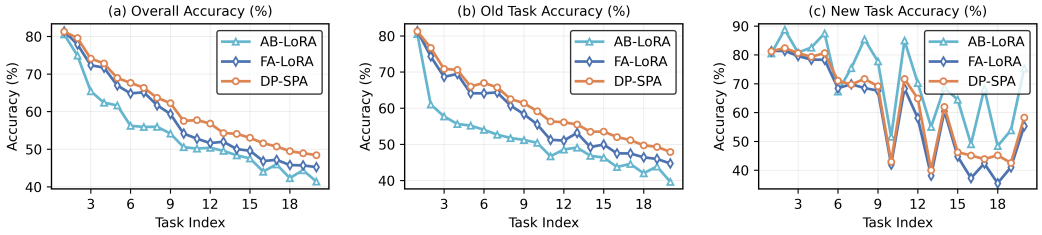

Figure 6: Overall, old-task, and new-task accuracy across 20 incremental tasks on ImageNet-A.