SEQUENTIAL ORTHOGONAL LOW-RANK ADAPTATION WITH FEATURE DISTRIBUTION DRIFT COMPENSATION FOR CONTINUAL LEARNING

Haibin Huang[1], **Jinglong Yang**[2], **Wenjian Huang**[1], **Jianguo Zhang**[1]
[1]Southern University of Science and Technology
[2]City University of Hong Kong

ABSTRACT

Orthogonal low-rank adaptation offers a parameter-efficient approach to class-incremental learning, but data-dependent subspace construction can introduce additional computational overhead. Moreover, orthogonality between parameter updates does not ensure stable feature representations, leaving stored historical class statistics misaligned with the evolving model and limiting the effectiveness of classifier alignment. To address these challenges, we propose Sequential Orthogonal Low-Rank Adaptation with Feature Distribution Drift Compensation. Our method sequentially allocates disjoint blocks of a shared, fixed orthonormal basis to the input-side LoRA factors, constructing mutually orthogonal adaptation subspaces without data-dependent subspace estimation. To account for the remaining feature drift, we fit a regularized residual transition using paired features extracted from current-task data before and after adaptation. This transition updates class means and covariances before classifier alignment, without retaining historical exemplars. By combining explicit parameter-space separation with feature-distribution correction, our framework addresses both subspace construction overhead and stale statistical representations. Results on standard and fine-grained class-incremental benchmarks show competitive accuracy in several settings, with dataset-dependent trade-offs, while simplifying orthogonal subspace construction.

1 INTRODUCTION

Class-incremental learning (CIL) requires a model to incorporate new classes sequentially while recognizing all previously observed classes without a task identifier. Updates based only on current data can overwrite information needed by earlier classes, motivating rehearsal, distillation, and regularization (Rebuffi et al., 2017; Li & Hoiem, 2018; Kirkpatrick et al., 2017). Pre-trained vision transformers (ViTs) provide a strong transferable representation (Dosovitskiy et al., 2021), but sequential adaptation still changes the feature space in which earlier decision boundaries were learned. In exemplar-free CIL, historical images are unavailable for directly correcting these changes.

Low-rank adaptation (LoRA) reduces the number of trainable parameters by adding a factorized residual to frozen weights (Hu et al., 2022). Continual variants separate successive updates through subspace constraints or decoupled parameterizations (Wang et al., 2023; Liang & Li, 2024; Wu et al., 2025). Data-dependent subspace estimation and orthogonality penalties, however, introduce additional computation. A fixed standard basis offers a simpler alternative, as explored by SBoRA and CSBoRA (Po et al., 2024; Liu et al., 2026): disjoint coordinate blocks yield orthogonal updates by construction. Yet orthogonal weight updates do not make old features invariant. A new branch can respond to old inputs, and changes in earlier transformer blocks propagate through subsequent nonlinear operations.

This distinction matters for classifier alignment (CA). Gaussian feature replay can calibrate a classifier across old and new classes without storing images (Zhang et al., 2023), but its historical means and covariances may describe obsolete representations. Aligning a classifier against stale distributions then creates a mismatch with the current feature extractor. Prior drift-compensation methods address prototype movement or distribution evolution (Yu et al., 2020; Gomez-Villa et al., 2024; Rypeść

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.

et al., 2024; Rao et al., 2026). Parameter-space separation and statistical correction therefore address different failure modes and should be considered together.

We develop a unified pipeline combining sequential fixed-basis LoRA with feature-distribution drift compensation. Only each task's output-side LoRA factors are learned. Paired current-task features before and after adaptation fit a regularized displacement model, which transports old class means. A separately regularized transition transports covariances and adds a shrunk within-class residual term. Balanced samples from the updated distributions then calibrate all classifier heads. The intended advantage is precise: no additional data-dependent orthogonal-space estimation is needed, while historical statistics are explicitly updated before CA. We do not claim zero total overhead or invariant representations.

Our contribution is the integration and analysis of these mechanisms for exemplar-free visual CIL, not the invention of standard-basis adaptation or drift compensation in isolation. We specify separate mean and covariance transport, expose their extrapolation assumptions, and examine their use with cumulative orthogonal adapters. Experiments cover four benchmarks, component combinations, and rank and ridge sensitivity. They support the usefulness of the complete alignment pipeline, but also reveal a Cars-196 accuracy deficit and the need to isolate transport from ordinary CA in further controlled experiments.

2 Related Work

Pre-trained models for class-incremental learning. L2P and DualPrompt learn prompts while preserving a pre-trained backbone (Wang et al., 2022b;a); CODA-Prompt uses decomposed attention-based prompts (Smith et al., 2023). SLCA instead couples slow representation adaptation with Gaussian classifier alignment (Zhang et al., 2023). These approaches offer different compromises between plasticity and stability. Our setting retains class statistics and historical model parameters, but neither old images nor per-example historical features, and uses a single cumulative predictor without task routing.

Continual low-rank adaptation. O-LoRA constrains low-rank subspaces in language-model continual learning (Wang et al., 2023); InfLoRA constructs interference-reduced adaptation subspaces for visual continual learning (Liang & Li, 2024). SD-LoRA decouples the scale and direction of low-rank updates (Wu et al., 2025). Most directly related to our parameterization, SBoRA fixes a standard-basis factor (Po et al., 2024), and CSBoRA allocates disjoint standard-basis directions across language tasks (Liu et al., 2026). We adopt this structural principle in a visual CIL pipeline and address the remaining distribution mismatch before CA. Thus, avoiding subspace estimation is an architectural property shared with this prior work; the empirical question here is how it interacts with explicit statistical transport.

Alignment and drift compensation. Semantic Drift Compensation estimates prototype movement from current samples (Yu et al., 2020). Learnable Drift Compensation learns a projector between successive representations (Gomez-Villa et al., 2024), while AdaGauss adapts covariances and addresses representation collapse (Rypeść et al., 2024). SLDC studies regularized transitions for distribution correction in sequentially fine-tuned ViTs (Rao et al., 2026). LCA improves classifier alignment through local loss stability (Tran et al., 2026). Our mean transition is a ridge-regularized residual model, and covariance transport uses a separately fitted map plus within-class residual compensation. The distinction from these methods lies in its coupling to cumulative fixed-basis adapters and the separation of the final moment maps, not a claim that old-class distribution drift has previously gone unaddressed.

3 Proposed Method

3.1 Setting and sequential orthogonal adaptation

At task $t \in \{1, \dots, T\}$, only $\mathcal{D}_t = \{(u_j, y_j)\}_{j=1}^{n_t}$ is available, with labels in $\mathcal{C}_t$ and disjoint class sets across tasks. Prediction ranges over $\mathcal{C}_{\le t} = \bigcup_{i \le t} \mathcal{C}_i$. Let $f_t(u) \in \mathbb{R}^{d_f}$ denote the adapted feature extractor. For each observed class c, memory retains a mean $\boldsymbol{\mu}_c$ and covariance $\boldsymbol{\Sigma}_c$, but no historical

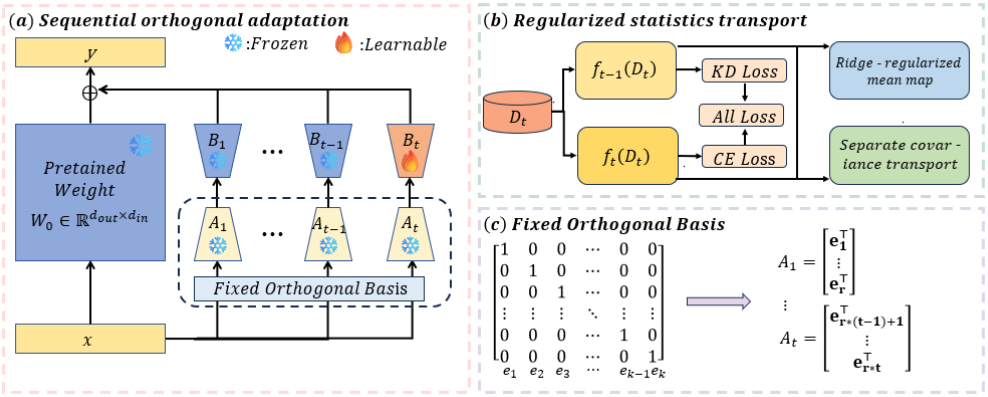

Figure 1: Overview of SOC-LoRA. (a) A frozen pretrained projection is augmented by sequential LoRA branches: the input factors $\mathbf{A}_i$ are fixed, disjoint basis blocks, while only the current output factor $\mathbf{B}_t$ is learned. (b) Paired features from $f_{t-1}(\mathcal{D}_t)$ and $f_t(\mathcal{D}_t)$ support KD/CE adaptation and separate ridge-regularized mean and covariance transport. (c) Each $\mathbf{A}_t$ selects a non-overlapping block from a shared canonical basis. Snowflakes and flames denote frozen and learnable components.

examples. Features are column vectors; superscripts $-$ and $+$ refer to states before and after the current adaptation.

For an adapted projection of input width d and rank r, task t receives a disjoint block of the standard basis:

$$\mathbf{U}_t = [\mathbf{e}_{(t-1)r+1}, \ldots, \mathbf{e}_{tr}], \quad \mathbf{A}_t = \mathbf{U}_t^\top, \quad \Delta\mathbf{W}_t = \mathbf{B}_t\mathbf{A}_t, \qquad Tr \le d. \tag{1}$$

Here $\mathbf{A}_t \in \mathbb{R}^{r \times d}$ is fixed and $\mathbf{B}_t \in \mathbb{R}^{d_{\text{out}} \times r}$ is zero-initialized and learned. Since $\mathbf{A}_i\mathbf{A}_j^\top = 0$ for $i \neq j$, cyclic invariance of the trace gives $\langle \Delta\mathbf{W}_i, \Delta\mathbf{W}_j \rangle_F = 0$, irrespective of the learned output factors. Subspace construction requires neither data-dependent decomposition nor an auxiliary orthogonality loss. In each adapted attention layer, key and value projections share the task's input directions but have separate output factors. Suppressing these indices,

$$\text{Proj}_t(\mathbf{x}) = \mathbf{W}_0\mathbf{x} + \mathbf{b}_0 + \sum_{i=1}^{t} \mathbf{B}_i\mathbf{A}_i\mathbf{x}. \tag{2}$$

Only the current $\mathbf{B}_t$ and current classifier head are updated; the backbone, input factors, historical adapters, and old heads are frozen. Historical branches remain active. The adaptation loss combines cross-entropy over current classes with old-class knowledge distillation (KD):

$$\mathcal{L}_{\text{adapt}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{KD}}\tau^2\mathbb{E}_{u \sim \mathcal{D}_t} \text{KL}\big(\mathbf{p}^-(u;\tau) \,\|\, \mathbf{p}^+(u;\tau)\big). \tag{3}$$

The distributions $\mathbf{p}^-$ and $\mathbf{p}^+$ are the teacher and student softmax outputs over $\mathcal{C}_{\le t-1}$ at temperature τ. KD is omitted at $t = 1$. Orthogonality of local updates does not imply $f_t(u) = f_{t-1}(u)$ on old images, motivating the next stage.

3.2 Regularized mean and covariance transport

For identical current images with matched deterministic preprocessing, extract $\mathbf{z}_j^- = f_{t-1}(u_j)$, $\mathbf{z}_j^+ = f_t(u_j)$, and displacement $\boldsymbol{\delta}_j = \mathbf{z}_j^+ - \mathbf{z}_j^-$. For $q \in \{\mu, \Sigma\}$, fit an affine residual map with a dedicated ridge coefficient:

$$(\mathbf{R}_q, \mathbf{b}_q) = \arg\min_{\mathbf{R},\mathbf{b}} \frac{1}{n_t} \sum_{j=1}^{n_t} \|\boldsymbol{\delta}_j - \mathbf{R}\mathbf{z}_j^- - \mathbf{b}\|_2^2 + \lambda_q\|\mathbf{R}\|_F^2, \qquad \mathbf{T}_q = \mathbf{I} + \mathbf{R}_q. \tag{4}$$

The unpenalized intercept preserves a shared translation. Every old mean is updated by

$$\boldsymbol{\mu}_c^+ = \mathbf{T}_\mu\boldsymbol{\mu}_c^- + \mathbf{b}_\mu = \boldsymbol{\mu}_c^- + \bar{\boldsymbol{\delta}} + \mathbf{R}_\mu(\boldsymbol{\mu}_c^- - \bar{\mathbf{z}}^-). \tag{5}$$

Bars denote averages over current samples. Increasing λ_μ suppresses position-dependent extrapolation; as $\lambda_\mu \to \infty$, the update tends to shared translation rather than leaving means unchanged. Applying this map to old classes assumes that drift learned on accessible classes transfers beyond its fitting distribution.

Covariances use the separately regularized map $\mathbf{T}_\Sigma$. Define residuals $\mathbf{e}_j = \mathbf{z}_j^+ - (\mathbf{T}_\Sigma \mathbf{z}_j^- + \mathbf{b}_\Sigma)$ and let $\mathbf{S}_{\text{res}}$ be the class-balanced average of their within-class sample covariances over $\mathcal{C}_t$. Historical covariance transport is

$$\mathbf{\Sigma}_c^+ = \mathbf{T}_\Sigma \mathbf{\Sigma}_c^- \mathbf{T}_\Sigma^\top + \beta\big[(1-\alpha)\mathbf{S}_{\text{res}} + \alpha \operatorname{diag}(\mathbf{S}_{\text{res}})\big] + \epsilon \mathbf{I}. \tag{6}$$

Here $\alpha \in [0,1]$ controls diagonal shrinkage, $\beta \geq 0$ scales residual compensation, and $\epsilon > 0$ stabilizes sampling. This estimator neglects input–residual cross-covariance and transfers current-class residual structure to old classes; it is not an exact identity for nonlinear feature evolution. Covariance-map parameters are selected on held-out current classes using moment-prediction error and then refitted. The final mean and covariance maps may differ; this does not require the covariance-parameter selector to ignore mean-prediction quality. No historical images or test data are needed by the proposed selection procedure.

3.3 Classifier alignment, cost, and scope

New-class moments are estimated directly from $\mathbf{z}_j^+$. For every seen class, sample M balanced pseudo-features $\widetilde{\mathbf{z}}_{c,m} \sim \mathcal{N}(\boldsymbol{\mu}_c^+, \mathbf{\Sigma}_c^+)$. Freeze f_t and optimize all classifier parameters $\boldsymbol{\phi}$ jointly:

$$\mathcal{L}_{\text{CA}} = -\frac{1}{M|\mathcal{C}_{\leq t}|} \sum_{c \in \mathcal{C}_{\leq t}} \sum_{m=1}^{M} \log p_{\boldsymbol{\phi}}(c \mid \widetilde{\mathbf{z}}_{c,m}). \tag{7}$$

Thus task-local training is followed by competition across the unified label space. Pseudo-features update only the classifier, not LoRA factors. At $t = 1$, historical transport and cross-task CA are omitted; otherwise statistics are recursively replaced after each task. Inference uses all accumulated adapters without a task identifier.

For square width-d key and value projections, the current trainable output factors contain $2dr$ parameters per adapted layer, excluding the classifier. Basis indices can represent fixed input factors without storing dense identity matrices. However, historical output factors accumulate, $Tr \leq d$ limits the number of disjoint directions, and full class covariances require $O(|\mathcal{C}_{\leq t}|d_f^2)$ storage. Ridge fitting, moment transport, and CA also have computational costs. Our claim concerns removing *data-dependent orthogonal-subspace construction*, not removing all overhead or guaranteeing zero forgetting.

4 EXPERIMENTS

4.1 EXPERIMENTAL PROTOCOL

Benchmarks and evaluation. We evaluate ten-stage class-incremental learning on CIFAR-100 (Krizhevsky, 2009), CUB-200-2011 (CUB-200) (Wah et al., 2011), Stanford Cars (Cars-196) (Krause et al., 2013), and ImageNet-R (Hendrycks et al., 2021). These benchmarks jointly cover generic recognition, fine-grained categorization, and distribution-shifted visual renditions. At stage t, the learner receives images only from the new class set $\mathcal{C}_t$ and is evaluated over all classes seen so far, $\mathcal{C}_{\leq t}$, without task identity. No historical images are retained.

Let $a_{t,k}$ denote accuracy on task k after learning task t, and let $w_{t,k}$ be its test-set weight, where $\sum_{k\leq t} w_{t,k} = 1$. We report

$$A_t = \sum_{k=1}^{t} w_{t,k} a_{t,k}, \qquad A_{\text{last}} = A_T, \qquad A_{\text{inc}} = \frac{1}{T}\sum_{t=1}^{T} A_t. \tag{8}$$

Last-Acc measures the quality of the final model, whereas Inc-Acc summarizes performance throughout the entire learning trajectory. All entries are percentages and are reported as mean $\pm$ standard deviation over three class orders.

Compared methods. The comparison spans four complementary families: sequential low-rank adaptation (Sequential LoRA (Hu et al., 2022), O-LoRA (Wang et al., 2023), InfLoRA (Liang & Li, 2024), and SD-LoRA (Wu et al., 2025)); prompt-based rehearsal-free learning (L2P (Wang et al., 2022b), DualPrompt (Wang et al., 2022a), and CODA-Prompt (Smith et al., 2023)); classifier-alignment methods (SLCA (Zhang et al., 2023) and β_1-SLDC (Rao et al., 2026)); and offline Joint Training. Joint Training has simultaneous access to all classes and is therefore shown only as a reference rather than a continual-learning competitor. SOC-LoRA uses the sequential fixed-basis adapters and drift-compensated classifier alignment described in Section 3.

Reporting scope. Tables 1 and 2 contain the author-provided aggregate outputs. This manuscript revision did not have access to per-order predictions or complete run logs; consequently, we report descriptive differences in percentage points (pp), rank methods by their means, and make no statistical-significance claim. The component study was produced as a separate controlled suite and its complete-configuration values differ slightly from the main comparison. We therefore compute every ablation delta strictly within Table 2.

Table 1: Main comparison under the common ten-stage class-incremental protocol (mean with SD in parentheses over three class orders, %). Bold marks the best incremental result including SOC-LoRA; underline marks the strongest competing result after excluding SOC-LoRA. Joint Training is an offline reference and is excluded from ranking.

Method	ImageNet-R		CIFAR-100		CUB-200		Cars-196	
	Last-Acc	Inc-Acc	Last-Acc	Inc-Acc	Last-Acc	Inc-Acc	Last-Acc	Inc-Acc
Joint Training	$82.76_{(0.54)}$	–	$93.13_{(0.21)}$	–	$88.26_{(0.73)}$	–	$80.31_{(0.13)}$	–
Sequential LoRA	$71.71_{(0.25)}$	$79.69_{(0.71)}$	$83.96_{(0.57)}$	$90.14_{(0.21)}$	$68.57_{(0.78)}$	$79.81_{(0.98)}$	$35.98_{(0.38)}$	$50.18_{(0.44)}$
O-LoRA	$73.30_{(0.47)}$	$80.59_{(0.66)}$	$85.39_{(0.49)}$	$91.12_{(0.35)}$	$68.08_{(2.07)}$	$79.23_{(1.56)}$	$40.84_{(1.48)}$	$55.73_{(0.33)}$
InfLoRA	$74.62_{(0.64)}$	$82.23_{(0.55)}$	$86.04_{(0.35)}$	$91.06_{(0.15)}$	$79.89_{(0.45)}$	$88.26_{(0.13)}$	$57.66_{(0.76)}$	$70.03_{(0.55)}$
L2P	$66.49_{(0.40)}$	$72.83_{(0.56)}$	$83.44_{(1.17)}$	$88.48_{(0.83)}$	$62.21_{(1.92)}$	$73.83_{(1.67)}$	$38.18_{(2.33)}$	$51.79_{(4.19)}$
DualPrompt	$66.07_{(0.53)}$	$71.91_{(0.27)}$	$83.87_{(0.21)}$	$89.45_{(0.73)}$	$73.82_{(0.57)}$	$84.61_{(0.50)}$	$40.08_{(2.36)}$	$56.01_{(1.78)}$
CODA-Prompt	$70.54_{(0.05)}$	$75.07_{(0.07)}$	$87.11_{(0.25)}$	$91.52_{(0.24)}$	$77.63_{(0.84)}$	$83.36_{(1.22)}$	$37.77_{(1.75)}$	$45.48_{(2.33)}$
SD-LoRA	$74.49_{(1.28)}$	$80.87_{(1.35)}$	$87.80_{(0.83)}$	$92.90_{(0.36)}$	$76.00_{(0.92)}$	$86.11_{(0.66)}$	$54.61_{(3.11)}$	$64.53_{(4.69)}$
SLCA	$77.15_{(0.21)}$	$81.17_{(0.64)}$	$89.66_{(0.34)}$	$93.78_{(0.55)}$	$83.88_{(0.44)}$	$89.55_{(0.68)}$	$73.09_{(1.16)}$	$79.87_{(2.34)}$
β_1-SLDC	$\underline{\mathbf{79.27}}_{(0.44)}$	$\underline{\mathbf{84.92}}_{(0.55)}$	$\underline{\mathbf{91.33}}_{(0.45)}$	$\underline{\mathbf{94.20}}_{(0.69)}$	$\underline{84.52}_{(1.41)}$	$\underline{90.62}_{(0.70)}$	$\underline{\mathbf{73.87}}_{(1.03)}$	$\underline{\mathbf{81.78}}_{(0.71)}$
SOC-LoRA (Ours)	$79.06_{(0.46)}$	$84.40_{(0.36)}$	$91.28_{(0.07)}$	$94.15_{(0.10)}$	$\mathbf{88.56}_{(0.33)}$	$\mathbf{92.24}_{(0.92)}$	$73.78_{(0.50)}$	$77.60_{(0.43)}$

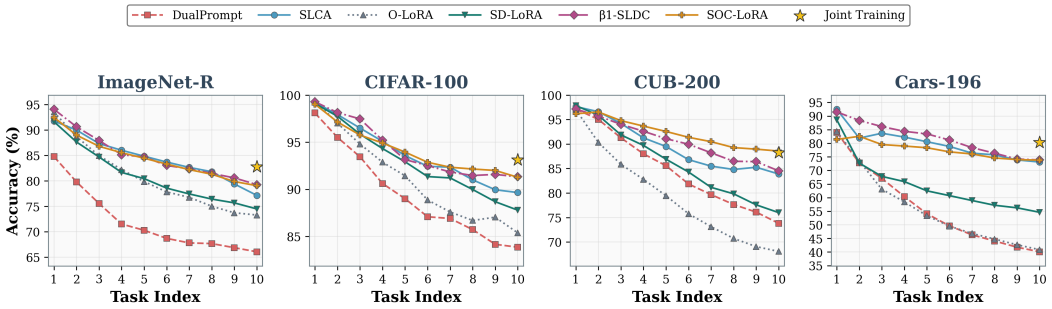

Figure 2: Accuracy over all classes observed after each of the ten incremental tasks. The curves complement the aggregate Last-Acc and Inc-Acc results in Table 1; Joint Training is shown only as an offline reference.

4.2 Main comparison

Strong performance across heterogeneous benchmarks. Table 1 shows that SOC-LoRA is consistently among the strongest incremental methods. On CUB-200, it obtains the best Last-Acc and Inc-Acc, 88.56% and 92.24%, improving over the second-best β_1-SLDC by 4.04 and 1.62 pp, respectively. The final result is also 0.30 pp above the reported Joint Training mean. Because Joint Training follows a different offline protocol and the reported standard deviations overlap, this observation should be read as closing the offline-reference gap, not as evidence that continual training is statistically superior.

On ImageNet-R, SOC-LoRA reaches $79.06\%/84.40\%$ Last-/Inc-Acc, only $0.21/0.52$ pp below β_1-SLDC and ahead of all remaining baselines. On CIFAR-100, the corresponding result is $91.28\%/94.15\%$, within 0.05 pp of the best reported mean for both metrics. Cars-196 presents a more nuanced trade-off: SOC-LoRA achieves the second-best final accuracy (73.78%, 0.09 pp below β_1-SLDC), while its 77.60% Inc-Acc trails β_1-SLDC and SLCA by 4.18 and 2.27 pp. Thus, the method finishes with a strong classifier on Cars-196 but does not dominate throughout the full trajectory.

Benefit over sequential low-rank adaptation. Relative to Sequential LoRA, SOC-LoRA improves Last-Acc by 7.35, 7.32, 19.99, and 37.80 pp on ImageNet-R, CIFAR-100, CUB-200, and Cars-196, respectively. The corresponding Inc-Acc gains are 4.71, 4.01, 12.43, and 27.42 pp. The largest improvements occur on the two fine-grained datasets, where small representation changes can substantially move class boundaries. SOC-LoRA also improves both metrics over O-LoRA and InfLoRA on every dataset in the main comparison. These results support the central design choice: exact separation of sequential low-rank update directions is useful, but high continual accuracy additionally requires correcting the feature statistics consumed by classifier alignment.

Across the four datasets, SOC-LoRA has a descriptive (unweighted) mean Last-Acc of 83.17%, compared with 82.25% for β_1-SLDC and 80.95% for SLCA. Its mean Inc-Acc is 87.10%, below β_1-SLDC (87.88%) but above SLCA (86.09%). This summary is not a pooled accuracy because the datasets have different class counts and test-set sizes; it is included only to characterize cross-dataset consistency.

Figure 2 exposes behavior that is hidden by a single final checkpoint. SOC-LoRA remains competitive throughout ImageNet-R and CIFAR-100 and approaches the offline reference on CUB-200. On Cars-196, its final accuracy is strong, but the trajectory remains below β_1-SLDC for most stages, explaining the larger gap in Inc-Acc. Because the evaluated label space grows at every stage, a decreasing all-seen accuracy curve is not by itself a pure measure of forgetting; task-by-stage accuracies would be required to compute backward transfer or average forgetting.

Table 2: Controlled component study (mean with SD in parentheses, %). KD denotes knowledge distillation and CA denotes drift-compensated classifier alignment. Bold marks the best incremental result including SOC-LoRA; underline marks the strongest competing result after excluding SOC-LoRA. Comparisons are made only within this separately run component suite.

Method	KD	CA	ImageNet-R		CIFAR-100		CUB-200		Cars-196	
			Last-Acc	Inc-Acc	Last-Acc	Inc-Acc	Last-Acc	Inc-Acc	Last-Acc	Inc-Acc
Joint Training	–	–	$82.76_{(0.54)}$	–	$93.13_{(0.21)}$	–	$88.26_{(0.73)}$	–	$80.31_{(0.13)}$	–
O-LoRA	–	–	$75.38_{(0.39)}$	$80.53_{(0.24)}$	$88.34_{(0.16)}$	$92.45_{(0.16)}$	$82.49_{(0.58)}$	$89.31_{(0.37)}$	$54.26_{(0.88)}$	$62.94_{(3.54)}$
O-LoRA	✓	–	$75.37_{(0.51)}$	$80.42_{(0.26)}$	$88.16_{(0.50)}$	$92.21_{(0.16)}$	$82.40_{(0.26)}$	$89.27_{(0.06)}$	$52.47_{(1.00)}$	$61.62_{(1.04)}$
O-LoRA	–	✓	$77.88_{(0.41)}$	$82.39_{(0.27)}$	**$90.78_{(0.28)}$**	$\underline{93.62}_{(0.27)}$	$88.89_{(0.75)}$	$92.52_{(0.35)}$	$65.67_{(0.42)}$	$72.73_{(0.49)}$
O-LoRA	✓	✓	$77.92_{(0.61)}$	$82.25_{(0.40)}$	$90.72_{(0.12)}$	$93.61_{(0.20)}$	**$\underline{89.18}_{(0.38)}$**	**$\underline{92.91}_{(0.68)}$**	$67.61_{(0.85)}$	$73.51_{(0.40)}$
InfLoRA	–	–	$77.59_{(0.52)}$	$83.45_{(0.31)}$	$85.20_{(0.31)}$	$90.57_{(0.25)}$	$79.89_{(0.45)}$	$88.26_{(0.12)}$	$57.66_{(0.65)}$	$70.03_{(0.45)}$
InfLoRA	✓	–	$77.44_{(0.37)}$	$84.10_{(0.22)}$	$85.57_{(0.09)}$	$91.15_{(0.18)}$	$80.99_{(0.60)}$	$88.90_{(0.41)}$	$63.32_{(0.92)}$	$72.63_{(0.80)}$
InfLoRA	–	✓	$\underline{78.00}_{(0.49)}$	$84.43_{(0.28)}$	$86.96_{(0.22)}$	$92.01_{(0.29)}$	$81.35_{(0.33)}$	$88.93_{(0.35)}$	$58.95_{(0.54)}$	$72.08_{(0.20)}$
InfLoRA	✓	✓	$77.87_{(0.61)}$	**$\underline{85.09}_{(0.35)}$**	$87.33_{(0.45)}$	$92.60_{(0.21)}$	$82.18_{(0.71)}$	$89.91_{(0.58)}$	$\underline{69.73}_{(0.45)}$	**$\underline{78.60}_{(0.17)}$**
SOC-LoRA (Ours)	–	–	$71.71_{(0.25)}$	$79.69_{(0.71)}$	$83.96_{(0.57)}$	$90.14_{(0.21)}$	$68.57_{(0.78)}$	$79.81_{(0.98)}$	$35.98_{(0.38)}$	$50.18_{(0.44)}$
SOC-LoRA (Ours)	✓	–	$73.32_{(0.65)}$	$80.36_{(0.48)}$	$85.56_{(0.13)}$	$90.95_{(0.23)}$	$69.20_{(0.78)}$	$80.22_{(1.12)}$	$40.03_{(1.31)}$	$52.74_{(0.29)}$
SOC-LoRA (Ours)	–	✓	$76.88_{(0.15)}$	$83.50_{(0.46)}$	$87.34_{(0.60)}$	$92.99_{(0.30)}$	$86.45_{(0.50)}$	$91.60_{(0.62)}$	$70.89_{(1.12)}$	$75.31_{(0.22)}$
SOC-LoRA (Ours)	✓	✓	**$79.18_{(0.08)}$**	$84.04_{(0.51)}$	$90.65_{(0.06)}$	**$94.01_{(0.11)}$**	$88.35_{(0.12)}$	$92.07_{(0.67)}$	**$72.47_{(0.55)}$**	$75.76_{(0.18)}$

4.3 Component analysis: KD and drift-aware CA

Classifier alignment is the dominant contributor. Table 2 evaluates knowledge distillation (KD) and drift-compensated classifier alignment (CA) on O-LoRA, InfLoRA, and SOC-LoRA. For SOC-LoRA, adding CA without KD raises Last-Acc over the adapter-only variant by 5.17, 3.38, 17.88, and 34.91 pp on ImageNet-R, CIFAR-100, CUB-200, and Cars-196. Inc-Acc increases by 3.81, 2.85, 11.79, and 25.13 pp. The much larger gains on CUB-200 and Cars-196 indicate that correcting the statistics used for classifier reconstruction is particularly important for fine-grained class boundaries.

The effect is not specific to our adapter. Without KD, enabling CA improves O-LoRA Last-Acc by 2.50, 2.44, 6.40, and 11.41 pp across the same datasets; for InfLoRA, the gains are 0.41, 1.76, 1.46, and 1.29 pp. The recurring improvements across distinct adapter parameterizations support feature-space calibration as a generally useful complement to parameter-space interference control.

KD and CA are complementary, but adapter dependent. For SOC-LoRA, KD alone provides smaller yet consistent Last-Acc gains of 1.61, 1.60, 0.63, and 4.05 pp. Starting from the CA-only variant, adding KD further improves Last-Acc by 2.30, 3.31, 1.90, and 1.58 pp, and Inc-Acc by 0.54, 1.02, 0.47, and 0.45 pp. Hence KD contributes beyond statistical alignment: the former constrains predictive change during representation learning, whereas the latter reconstructs a balanced classifier in the updated feature space.

The interaction is not universally additive. KD without CA is neutral or slightly detrimental for O-LoRA in this suite, whereas the complete InfLoRA configuration produces a particularly large Cars-196 improvement. Under the complete setting, SOC-LoRA attains higher final accuracy than both O-LoRA and InfLoRA on ImageNet-R and Cars-196, is within 0.07 pp of O-LoRA on CIFAR-100, and is 0.83 pp below O-LoRA on CUB-200. This pattern argues against attributing the gains to a single generic post-processing step: the adapter geometry, distillation objective, and transported class statistics interact in a dataset-dependent manner.

4.4 Discussion and limitations

The results support two conclusions. First, the fixed-basis construction remains competitive with learned or data-dependent adaptation spaces while avoiding a separate basis-estimation procedure. Second, the component study shows that classifier alignment based on updated feature distributions is essential to the complete system, especially on fine-grained datasets. These are complementary advantages: orthogonal parameter updates reduce direct overlap, but they do not by themselves keep historical class moments valid after the representation changes.

The current evidence also defines the boundary of the claim. Table 2 compares CA on/off, where the enabled condition already contains drift compensation; it does not include a stale-statistics CA control. The table therefore validates the complete drift-aware alignment pipeline but does not isolate how much of its gain comes from distribution transport versus class-balanced classifier refitting. A decisive follow-up should compare, under identical seeds and sampling budgets, (i) no CA, (ii) ordinary CA with uncorrected moments, (iii) mean-only transport, (iv) mean-and-covariance transport, and (v) the full residual-compensation model. Likewise, the decomposition-free basis construction is an architectural fact, while its wall-clock advantage requires matched runtime and peak-memory measurements. Reporting these controls together with paired per-order results would turn the present aggregate evidence into a direct test of both proposed advantages.

5 CONCLUSION

We presented SOC-LoRA, an exemplar-free CIL pipeline that combines sequential fixed-basis orthogonal adaptation with feature-distribution drift compensation. Disjoint input directions eliminate a separate data-dependent orthogonal-space estimation step, while transported class moments reduce the mismatch between historical statistics and the current representation before classifier alignment. Across four benchmarks, SOC-LoRA achieves the best reported CUB-200 result, top-two final accuracy on the other three datasets, and large gains over sequential LoRA. The controlled component study identifies drift-aware classifier alignment as the dominant contributor on fine-grained data and shows that KD provides a complementary benefit for SOC-LoRA. The results also reveal a Cars-196 trajectory trade-off and adapter-dependent KD/CA interactions. A stale-statistics CA control, paired per-order tests, and matched runtime and memory measurements remain necessary to isolate the transport mechanism and quantify the end-to-end efficiency advantage.

REFERENCES

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.

Alex Gomez-Villa, Dipam Goswami, Kai Wang, Andrew D. Bagdanov, Bartlomiej Twardowski, and Joost van de Weijer. Exemplar-free continual representation learning via learnable drift compensation. In *European Conference on Computer Vision*, pp. 473–490, 2024. doi: 10.1007/978-3-031-72667-5_27.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021. URL https://arxiv.org/abs/2006.16241.

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.

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

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013. doi: 10.1109/ICCVW.2013.77.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018.

Yan-Shuo Liang and Wu-Jun Li. InfLoRA: Interference-free low-rank adaptation for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. doi: 10.1109/CVPR52733.2024.02231.

Yuyang Liu, Lai-Man Po, Farrell Hung, Zhuohan Wang, Haoxuan Wu, Zeyu Jiang, Kun Li, Xuyuan Xu, and Kwok-Wai Cheung. CSBoRA: A continual learning method for large language models with true orthogonality and reduced forgetting. *Pattern Recognition*, 179:113782, 2026. doi: 10.1016/j.patcog.2026.113782.

Lai-Man Po, Yuyang Liu, Haoxuan Wu, Tianqi Zhang, Wing-Yin Yu, Zhuohan Wang, Zeyu Jiang, and Kun Li. SBoRA: Low-rank adaptation with regional weight updates. *arXiv preprint arXiv:2407.05413*, 2024. URL https://arxiv.org/abs/2407.05413.

Xuan Rao, Simian Xu, Zheng Li, Bo Zhao, Derong Liu, Mingming Ha, and Cesare Alippi. Compensating distribution drifts in continual learning with pre-trained vision transformers. In *AAAI Conference on Artificial Intelligence*, 2026.

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

Grzegorz Rypeść, Sebastian Cygert, Tomasz Trzcinski, and Bartlomiej Twardowski. Task-recency bias strikes back: Adapting covariances in exemplar-free class-incremental learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

James Seale II Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. CODA-Prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Tung Tran, Danilo Vasconcellos Vargas, and Khoat Than. LCA: Local classifier alignment for continual learning. In *International Conference on Learning Representations*, 2026.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. URL `https://www.vision.caltech.edu/datasets/cub_200_2011/`.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10658–10671, 2023. doi: 10.18653/v1/2023.findings-emnlp.715.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, pp. 631–648, 2022a.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 139–149, 2022b.

Yichen Wu, Hongming Piao, Long-Kai Huang, Renzhen Wang, Wanhua Li, Hanspeter Pfister, Deyu Meng, Kede Ma, and Ying Wei. SD-LoRA: Scalable decoupled low-rank adaptation for class-incremental learning. In *International Conference on Learning Representations*, 2025.

Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6982–6991, 2020.

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

A SUPPLEMENTARY EXPERIMENTAL SUMMARIES

This supplement preserves the scope of the supplied exports. Source tables are retained unchanged under `tables/source/`; vector plots under `figures/experiments/`. The main comparison is retypeset in Table 1. No numbers are perturbed, simulated, or substituted from the example screenshot. The following tables reorganize the remaining results for readability. Source captions describe three class orders, but per-order logs and seed identities were not supplied. An unweighted dataset mean is descriptive and is not pooled test accuracy.

Reproducibility and AI assistance statement. No new training runs were performed during manuscript preparation. Missing metadata and unresolved aggregation issues are identified in Section 4. Generative AI assisted with drafting, literature organization, arithmetic checks, and typesetting, not with generating experimental measurements. The authors must verify the underlying runs, implementation, scientific claims, and references before submission.

A.1 COMPONENT-STUDY PROVENANCE

Table 2 in the main paper retains all reported adapter/KD/CA combinations. The component study is kept separate from the main comparison because the complete-configuration values differ slightly across the two supplied exports. Matching optimization, parameter counts, and the exact CA implementation across adapters must be verified from the run configurations before submission. Because the enabled CA condition includes drift compensation, it is not an isolated comparison of transport against uncorrected CA.

A.2 COMPLETE COMPONENT METRICS

Table 3 includes incremental accuracy as well as the final accuracy reported in the main text. Figure 3 is the corresponding supplied bar chart. It visualizes the combined operation, not an independent transport ablation.

Table 3: Complete component summaries (mean ± SD, %). Inc and Last abbreviate A_{avg} and A_{last}.

Configuration	Metric	CIFAR-100	CUB-200	Cars-196	ImageNet-R
Adaptation only	Inc	$90.14_{\pm 0.21}$	$79.81_{\pm 0.98}$	$50.18_{\pm 0.44}$	$79.69_{\pm 0.71}$
	Last	$83.96_{\pm 0.57}$	$68.57_{\pm 0.78}$	$35.98_{\pm 0.38}$	$71.71_{\pm 0.25}$
+ KD	Inc	$90.95_{\pm 0.23}$	$80.22_{\pm 1.12}$	$52.74_{\pm 0.29}$	$80.36_{\pm 0.48}$
	Last	$85.56_{\pm 0.13}$	$69.20_{\pm 0.78}$	$40.03_{\pm 1.31}$	$73.32_{\pm 0.65}$
+ CA†	Inc	$92.99_{\pm 0.30}$	$91.60_{\pm 0.62}$	$75.31_{\pm 0.22}$	$83.50_{\pm 0.46}$
	Last	$87.34_{\pm 0.60}$	$86.45_{\pm 0.50}$	$70.89_{\pm 1.12}$	$76.88_{\pm 0.15}$
+ KD + CA†	Inc	$94.01_{\pm 0.11}$	$92.07_{\pm 0.67}$	$75.76_{\pm 0.18}$	$84.04_{\pm 0.51}$
	Last	$90.65_{\pm 0.06}$	$88.35_{\pm 0.12}$	$72.47_{\pm 0.55}$	$79.18_{\pm 0.08}$

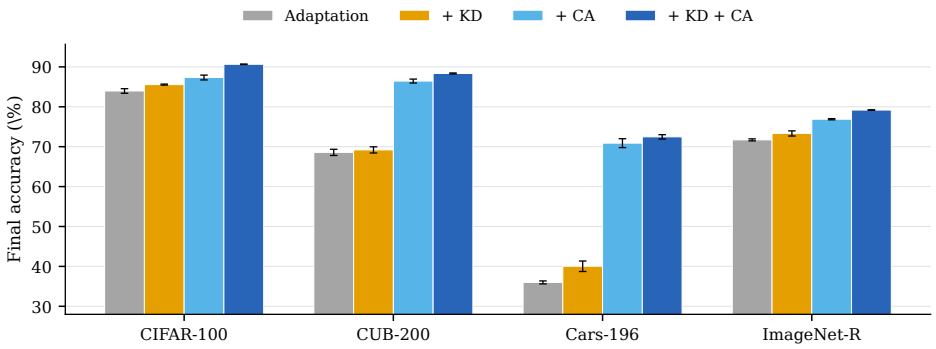

Figure 3: Supplied component chart. Its "CA" label corresponds to CA† in the manuscript: classifier alignment with drift-compensated distributions. Error bars and averaging conventions must be checked against the underlying plotting code.

A.3 RANK AND RIDGE: EXACT EXPORTED VALUES

Tables 4 and 5 contain all values from the supplied sweeps. The ridge export states that training budgets are held constant and settings are not selected on test data; the corresponding selection logs were not available for verification. Values are not evidence of statistical significance. In particular, the rank figure additionally contains incremental-accuracy curves for which no tabulated values were supplied; these have not been reverse-engineered or invented.

Table 4: Sensitivity to the per-task LoRA rank with $\lambda_\mu = 10$. Entries are final-stage accuracy over three class orders.

Rank	CIFAR-100 A_{last}	CUB-200 A_{last}	Cars-196 A_{last}	ImageNet-R A_{last}	Mean
5	91.28 ± 0.07	88.56 ± 0.33	70.12 ± 0.32	79.02 ± 0.13	82.25
8	91.01 ± 0.27	88.45 ± 0.23	71.63 ± 0.43	79.27 ± 0.12	82.59
10	90.65 ± 0.06	88.35 ± 0.12	72.47 ± 0.55	79.18 ± 0.08	82.66
16	90.46 ± 0.17	87.12 ± 0.64	72.69 ± 0.18	79.06 ± 0.46	82.33
20	90.20 ± 0.29	85.87 ± 0.28	73.78 ± 0.50	78.85 ± 0.53	82.17

Table 5: Mean-transport ridge sensitivity on the fine-grained benchmarks (mean $\pm$ SD, %). The export states a fixed training budget and no test-set selection; the corresponding run metadata remains to be verified.

λ_μ	CUB-200 A_{avg}	CUB-200 A_{last}	Cars-196 A_{avg}	Cars-196 A_{last}
0.01	89.78 ± 0.76	83.53 ± 1.18	73.66 ± 0.55	68.87 ± 0.95
1	91.95 ± 0.68	87.87 ± 0.17	75.94 ± 0.27	72.76 ± 0.64
2	92.06 ± 0.67	88.23 ± 0.19	75.92 ± 0.26	72.69 ± 0.60
3	92.07 ± 0.64	88.26 ± 0.19	75.87 ± 0.24	72.75 ± 0.62
5	92.09 ± 0.70	88.35 ± 0.19	75.82 ± 0.24	72.53 ± 0.63
10	92.07 ± 0.67	88.35 ± 0.12	75.76 ± 0.18	72.47 ± 0.55
15	92.04 ± 0.67	88.32 ± 0.22	75.76 ± 0.16	72.55 ± 0.51
30	92.00 ± 0.72	88.20 ± 0.23	75.50 ± 0.14	72.10 ± 0.44

A.4 OLD VERSUS NEWEST-CLASS ACCURACY

Table 6 and Figure 4 preserve the old/new diagnostics. For each individual run with sample-weighted accuracy, the identity $A_T = (1-\rho)A_{\text{old}} + \rho A_{\text{new}}$ must hold, where ρ is the fraction of newest-task test samples. Macro-averaging requires corresponding class weights instead. With $\rho = 0.1$, the proposed ImageNet-R old/new values give 78.165%, rather than the exported 79.18%. This is an aggregation check, not proof of an incorrect result: actual task counts, weighting, and per-run pairing must be obtained. The ImageNet-R decomposition is therefore provisional and excluded from quantitative retention conclusions. If weights vary across runs, the identity must be checked before averaging rather than imposed on marginal means.

Table 6: Final-stage old/new diagnostics (mean ± SD, %). New refers only to the newest task. ImageNet-R values are provisional pending aggregation checks.

Method	Group	CIFAR-100	CUB-200	Cars-196	ImageNet-R
Normal LoRA	Old	$90.27_{\pm 0.43}$	$85.78_{\pm 0.30}$	$74.90_{\pm 0.67}$	$77.67_{\pm 0.67}$
	New	$93.27_{\pm 0.65}$	$94.60_{\pm 1.43}$	$81.34_{\pm 2.79}$	$83.56_{\pm 3.42}$
O-LoRA	Old	$90.80_{\pm 0.18}$	$88.34_{\pm 0.31}$	$74.34_{\pm 0.83}$	$78.77_{\pm 0.23}$
	New	$92.93_{\pm 0.99}$	$93.25_{\pm 1.73}$	$79.35_{\pm 2.44}$	$82.65_{\pm 3.03}$
InfLoRA	Old	$90.56_{\pm 0.25}$	$86.11_{\pm 0.59}$	$73.23_{\pm 0.54}$	$78.17_{\pm 0.49}$
	New	$92.07_{\pm 0.47}$	$94.13_{\pm 0.93}$	$78.63_{\pm 2.45}$	$81.49_{\pm 3.27}$
Ours	Old	$90.39_{\pm 0.19}$	$87.87_{\pm 0.27}$	$71.98_{\pm 0.58}$	$77.67_{\pm 0.05}$
	New	$93.00_{\pm 1.39}$	$93.52_{\pm 0.94}$	$76.70_{\pm 2.67}$	$82.62_{\pm 3.57}$

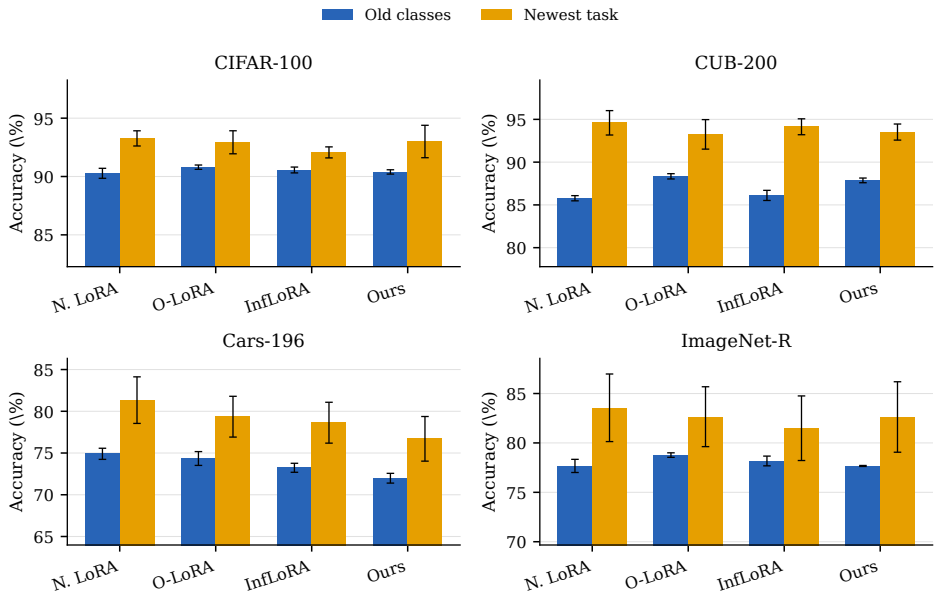

Figure 4: Supplied old/new diagnostic, retained for traceability. ImageNet-R aggregation remains unresolved; see the text above. The other datasets also require per-run weights and test counts for a full consistency audit.