BEYOND PARAMETER PROTECTION: FISHER-SHAPED EWC FOR REPLAY-FREE CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In replay-free continual learning, Elastic Weight Consolidation (EWC) uses Fisher information matrix (FIM) to protect acquired knowledge by penalizing changes to historically important parameters. However, a local importance estimate at the retained solution does not in general characterize sensitivity after finite parameter displacement. This motivates controlling Fisher sensitivity over a neighborhood of potential future updates. We propose Fisher-Shaped EWC (FS-EWC), which extends the role of FIM from parameter protection to solution formation for future retention. During task learning, FS-EWC minimizes the worst-case empirical FIM trace within a prescribed neighborhood, explicitly favoring solutions that are less sensitive to bounded future displacement. At task completion, trajectory-based model merging constructs the retained solution from multiple optimization states rather than a single terminal iterate, after which FIM is estimated for subsequent consolidation. Experiments across multiple benchmarks show that FS-EWC improves performance over EWC, while controlled studies indicate a more favorable stability–plasticity balance and greater tolerance to parameter drift. These results suggest that effective knowledge retention requires both forming solutions with future retention in mind and protecting acquired knowledge.

1 INTRODUCTION

Continual learning (CL) aims to acquire new knowledge over time while retaining previously learned capabilities, which is essential in real-world environments where data distributions and learning demands continuously evolve (Wang et al., 2024; Shaheen et al., 2022). This challenge is particularly pronounced in replay-free CL, where historical data are unavailable during subsequent training. Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017) is a representative and widely used regularization baseline for mitigating forgetting in this setting. However, it still faces a persistent stability–plasticity trade-off between retaining acquired knowledge and adapting to new tasks.

EWC constructs future protection from the learned solution and its fisher information matrix (FIM), penalizing changes to parameters that are important to previously learned tasks. After each task, the resulting solution is retained as the reference for subsequent learning, while FIM is consolidated with historical parameter importance (Huszár, 2018; Schwarz et al., 2018). Future updates are penalized according to this inherited importance, only allowing movement along directions assigned lower FIM values. However, this protection is based on a local characterization of parameter importance. As subsequent learning moves the model away from the learned solution, directions that initially appear less important can become increasingly sensitive. This motivates learning a solution whose surrounding region is less susceptible to future parameter drift, rather than relying only on constraints imposed during later task.

Existing extensions of EWC mainly strengthen how parameter importance is constructed and used, while still treating the solution to be protected as given. They do not explicitly shape the current solution so that its surrounding neighborhood remains less sensitive to future parameter drift. Online and trajectory-aware approaches refine when and how importance is accumulated across learning (Huszár, 2018; Schwarz et al., 2018; Zenke et al., 2017; Chaudhry et al., 2018); studies of FIM computation examine how the FIM is computed (Martens, 2020; van de Ven, 2025); and structured approximations improve how this information is represented beyond a diagonal form (Martens & Grosse, 2015; Liu et al., 2018).

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.

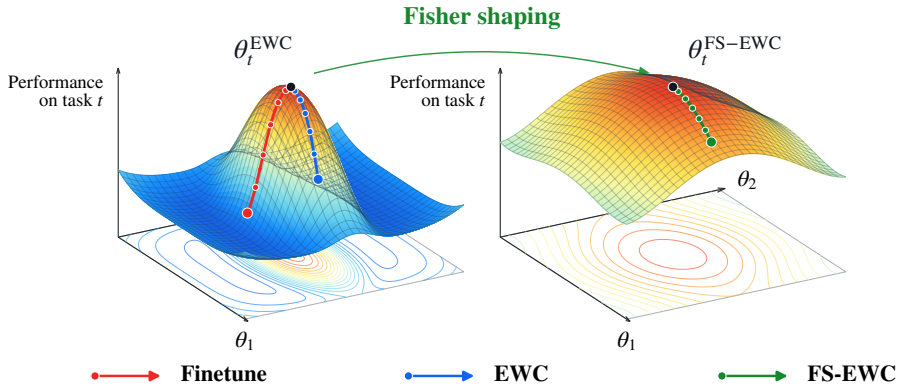


Figure 1: **Beyond protection: shaping solutions for future retention.** The surfaces schematically depict task-t performance around learned solutions, while trajectories illustrate parameter updates during task-$t+1$ learning. Unlike finetuning, EWC penalizes subsequent drift using historical parameter importance. FS-EWC additionally penalizes the worst-case empirical FIM over a prescribed neighborhood during task-t learning, favoring solutions less sensitive to bounded future drift while preserving historical protection. Complemented by model merging over task trajectory, our approach incorporates future retention into how solutions are learned and retained for subsequent protection.

The missing component is an explicit criterion for forming solutions that remain less sensitive to future parameter displacement. We derive a finite-displacement bound that relates task-specific loss changes to FIM along the displacement path. For a prescribed displacement radius, reducing the worst-case FIM trace over the surrounding neighborhood tightens the bound on potential loss increase. This result makes future retention an explicit consideration during current-task learning and leads to our central design principle: FIM should guide solution formation as well as subsequent parameter protection.

We instantiate this principle with **Fisher-Shaped EWC (FS-EWC)**, which shapes both how the current solution is learned and how the retained reference is constructed. As illustrated in Fig. 1, EWC constrains subsequent parameter drift using historical importance, whereas FS-EWC additionally shapes the solution before it is inherited by future tasks. During current-task learning, FS-EWC augments the EWC objective with a neighborhood FIM-trace penalty, making sensitivity to bounded future displacement an explicit optimization criterion alongside task fitting and historical regularization. At task completion, trajectory-based model merging constructs the retained reference from multiple optimization states rather than directly committing a single terminal iterate. FIM is then estimated at the retained solution and consolidated for subsequent protection. The two components address complementary stages of the same process: shaping the solutions produced during learning and constructing the solution passed forward for future.

Experiments across diverse vision and language benchmarks show that FS-EWC generally improves upon the corresponding EWC baselines under both full-parameter and LoRA. Controlled studies further show that shaping the solution itself improves subsequent retention under the same protection rule, demonstrating that future forgetting depends not only on how a solution is protected but also on the solution being protected. When combined with historical protection, solution shaping improves both retained-task and new-task accuracy in the controlled comparisons. The gains also persist across alternative FIM computation and representations. Together, these results suggest that the stability–plasticity balance can be improved by jointly shaping solutions for future retention and protecting acquired knowledge during subsequent learning.

Our contributions are summarized as follows:

- We identify a limitation of EWC: it protects a learned solution using locally estimated parameter importance, but does not explicitly account for how sensitive that solution may become under finite future parameter drift. By analyzing task-loss changes along such displacement, we show that subsequent retention depends on Fisher sensitivity over the surrounding parameter neighborhood, motivating the solution itself to be shaped for future.

- We propose Fisher-Shaped EWC (FS-EWC), which extends EWC from protecting learned solutions to explicitly shaping and retaining solutions for future tasks. During task learning, FS-EWC reduces neighborhood FIM sensitivity while preserving historical EWC regularization. At task completion, trajectory-based model merging constructs the retained anchor, at which current-task FIM is estimated for subsequent protection.
- We propose a broader design principle for continual learning:: the stability–plasticity trade-off can be mitigated by incorporating future retention into solution formation alongside historical knowledge protection. Experiments across vision and language benchmarks support this principle, showing gains over corresponding EWC baselines and complementarity with alternative FIM computation schemes and matrix representations.

2 Related Works

Continual Learning. Continual learning seeks to preserve previously acquired knowledge while enabling adaptation to evolving tasks and data distributions (Wang et al., 2024). A central challenge is the stability–plasticity trade-off: restricting model changes can preserve previous knowledge but may limit adaptation, whereas greater plasticity can accelerate forgetting. Replay-based methods alleviate this tension by revisiting previous examples, while replay-free learning must balance retention and adaptation without access to historical data. We focus on replay-free weight regularization, which addresses this problem by constraining parameter updates during sequential learning.

Elastic Weight Consolidation. Elastic Weight Consolidation (EWC) is a widely used regularization baseline for replay-free continual learning (Kirkpatrick et al., 2017). Subsequent work has strengthened this protection mechanism by refining how historical importance is accumulated and used across sequential learning. Online consolidation methods recursively update inherited importance across tasks (Huszár, 2018; Schwarz et al., 2018), while trajectory-aware approaches estimate parameter importance from the optimization process (Zenke et al., 2017; Chaudhry et al., 2018). More recently, EWC-style regularization has been extended to low-rank continual learning, demonstrating its applicability to LoRA-based adaptation (Zheng et al., 2026). These approaches improve how acquired knowledge is protected during subsequent learning, but largely treat the solution being protected as given.

FIM Estimation and Representation. The effectiveness of FIM-based regularization method also depends on how FIM is computed and represented. Prior studies examine the distinction between empirical and model-based FIM and the consequences of practical computational approximations for continual learning (Martens, 2020; van de Ven, 2025). Other work improves the representation of FIM beyond diagonal approximations, including rotated and Kronecker-factored structures (Liu et al., 2018; Martens & Grosse, 2015). These advances improve the fidelity or expressiveness of the importance signal used to protect a given reference solution. Our work addresses an orthogonal dimension: how the reference solution itself can be formed to better support future task.

3 Methodology

3.1 Preliminaries

Problem formulation. We consider replay-free class-incremental learning with tasks $\mathbb{T}_1, \ldots, \mathbb{T}_T$. At task t, the learner receives $\mathbb{S}_t = \{(\boldsymbol{x}_i^{(t)}, y_i^{(t)})\}_{i=1}^{m_t}$ sampled from $\mathcal{D}_t$. Inference distinguishes all observed classes without task identities. Training accesses only $\mathbb{S}_t$; historical examples are neither retained nor replayed. Let $f_{\boldsymbol{\theta}}$ have trainable parameters $\boldsymbol{\theta} \in \mathbb{R}^d$ and predictive probabilities $p_{\boldsymbol{\theta}}(y \mid \boldsymbol{x})$. Define

$$\ell_{t,i}^{\mathrm{CE}}(\boldsymbol{\theta}) = -\log p_{\boldsymbol{\theta}}(y_i^{(t)} \mid \boldsymbol{x}_i^{(t)}), \qquad \widehat{\mathcal{L}}_t^{\mathrm{CE}}(\boldsymbol{\theta}) = \frac{1}{m_t} \sum_{i=1}^{m_t} \ell_{t,i}^{\mathrm{CE}}(\boldsymbol{\theta}). \tag{1}$$

The parameters retained after task t are denoted by $\boldsymbol{\theta}_t$.

EWC and empirical FIM. EWC (Kirkpatrick et al., 2017) retains a reference solution together with a FIM-based summary of parameter importance, and penalizes subsequent changes along directions that are important to previously learned tasks. We adopt the online formulation (Huszár,

2018; Schwarz et al., 2018):

$$\mathcal{J}_t^{\mathrm{EWC}}(\boldsymbol{\theta}) = \widehat{\mathcal{L}}_t^{\mathrm{CE}}(\boldsymbol{\theta}) + \Omega_{t-1}(\boldsymbol{\theta}), \quad \Omega_{t-1}(\boldsymbol{\theta}) = \frac{\lambda}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1})^\top \boldsymbol{\Lambda}_{t-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}). \tag{2}$$

Here, $\boldsymbol{\theta}_{t-1}$ is the retained reference from the preceding task, and $\boldsymbol{\Lambda}_{t-1} \succeq \mathbf{0}$ denotes the consolidated diagonal importance matrix. Both are fixed during task-t optimization, while $\lambda \geq 0$ controls the strength of historical protection.

For the current task, let $\boldsymbol{g}_{t,i}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \ell_{t,i}^{\mathrm{CE}}(\boldsymbol{\theta})$. Define the empirical FIM as

$$\boldsymbol{F}_t^{\mathrm{emp}}(\boldsymbol{\theta}) = \frac{1}{m_t} \sum_{i=1}^{m_t} \boldsymbol{g}_{t,i}(\boldsymbol{\theta}) \boldsymbol{g}_{t,i}(\boldsymbol{\theta})^\top. \tag{3}$$

Its diagonal provides the task-specific importance contribution used for subsequent EWC consolidation, while its trace will also serve below as a scalar measure of local solution sensitivity.

3.2 Solution Sensitivity to Future Drift

The importance stored at an anchor does not by itself determine its tolerance to finite future drift. For a candidate task-t solution $\boldsymbol{\theta}$, define the loss change induced by displacement $\boldsymbol{\Delta}$ as

$$D_t(\boldsymbol{\theta}, \boldsymbol{\Delta}) = \widehat{\mathcal{L}}_t^{\mathrm{CE}}(\boldsymbol{\theta} + \boldsymbol{\Delta}) - \widehat{\mathcal{L}}_t^{\mathrm{CE}}(\boldsymbol{\theta}). \tag{4}$$

Positive values describe deterioration in task-t empirical loss. For a later retained model $\boldsymbol{\theta}_k$, $k > t$, the corresponding displacement is $\boldsymbol{\theta}_k - \boldsymbol{\theta}_t$. Since this displacement is unknown during task-t learning, we consider the prescribed neighborhood $\mathbb{U}_r(\boldsymbol{\theta}) = \{\boldsymbol{\theta} + \boldsymbol{\epsilon} : \|\boldsymbol{\epsilon}\|_2 \leq r\}$.

Proposition 1 (Finite-displacement sensitivity) *Suppose each $\ell_{t,i}^{\mathrm{CE}}$ is continuously differentiable on an open set containing $\mathbb{U}_r(\boldsymbol{\theta})$. Then every $\|\boldsymbol{\Delta}\|_2 \leq r$ satisfies*

$$|D_t(\boldsymbol{\theta}, \boldsymbol{\Delta})|^2 \leq \int_0^1 \boldsymbol{\Delta}^\top \boldsymbol{F}_t^{\mathrm{emp}}(\boldsymbol{\theta} + \tau\boldsymbol{\Delta}) \boldsymbol{\Delta} \, \mathrm{d}\tau. \tag{5}$$

The proof is in Appendix A.3.

To obtain a scalar criterion without specifying a displacement direction, observe that

$$\boldsymbol{\Delta}^\top \boldsymbol{F}_t^{\mathrm{emp}}(\boldsymbol{u}) \boldsymbol{\Delta} \leq \|\boldsymbol{\Delta}\|_2^2 \frac{1}{m_t} \sum_{i=1}^{m_t} \|\boldsymbol{g}_{t,i}(\boldsymbol{u})\|_2^2. \tag{6}$$

Accordingly, define the point FIM sensitivity

$$\mathcal{R}_{\mathrm{F},t}(\boldsymbol{\theta}) = \frac{1}{m_t} \sum_{i=1}^{m_t} \|\boldsymbol{g}_{t,i}(\boldsymbol{\theta})\|_2^2 = \operatorname{tr} \boldsymbol{F}_t^{\mathrm{emp}}(\boldsymbol{\theta}). \tag{7}$$

Combining these expressions yields

$$|D_t(\boldsymbol{\theta}, \boldsymbol{\Delta})|^2 \leq \|\boldsymbol{\Delta}\|_2^2 \int_0^1 \mathcal{R}_{\mathrm{F},t}(\boldsymbol{\theta} + \tau\boldsymbol{\Delta}) \, \mathrm{d}\tau. \tag{8}$$

This bound shows that finite-displacement loss change depends on FIM sensitivity along the displacement path, rather than only at the anchor. Since the future path is unknown during task-t learning, this motivates controlling sensitivity over a bounded neighborhood around the current solution, rather than relying only on importance estimated at the retained anchor.

3.3 Fisher-Shaped EWC

FS-EWC combines neighborhood sensitivity shaping with an explicit rule for constructing the retained anchor. The former guides current-task learning while preserving the inherited EWC penalty; the latter selects the parameters at which current-task importance is estimated for subsequent learning. A complementary distributional interpretation is provided in Appendix A.4, while the complete FS-EWC algorithm are detailed in Appendix A.5.

3.3.1 NEIGHBORHOOD FISHER SENSITIVITY SHAPING

Over the prescribed neighborhood of radius r, we define the worst-case FIM sensitivity as

$$\mathcal{R}_{\mathrm{F},t}^{\max}(\boldsymbol{\theta}; r) = \max_{\|\boldsymbol{\epsilon}\|_2 \le r} \mathcal{R}_{\mathrm{F},t}(\boldsymbol{\theta} + \boldsymbol{\epsilon}) = \max_{\|\boldsymbol{\epsilon}\|_2 \le r} \operatorname{tr} \boldsymbol{F}_t^{\mathrm{emp}}(\boldsymbol{\theta} + \boldsymbol{\epsilon}). \tag{9}$$

Here, $\boldsymbol{\epsilon}$ specifies locations at which sensitivity is evaluated, whereas $\boldsymbol{\Delta}$ denotes possible future parameter drift. The maximization characterizes the largest FIM sensitivity within this neighborhood, without assuming a particular direction or distribution of future drift.

Combining Eq. equation 8 with Young's inequality gives

$$\sup_{\|\boldsymbol{\Delta}\|_2 \le r} |D_t(\boldsymbol{\theta}, \boldsymbol{\Delta})| \le r\sqrt{\mathcal{R}_{\mathrm{F},t}^{\max}(\boldsymbol{\theta}; r)} \le \mathcal{R}_{\mathrm{F},t}^{\max}(\boldsymbol{\theta}; r) + \frac{r^2}{4}. \tag{10}$$

Since the final term is independent of $\boldsymbol{\theta}$, this motivates

$$\mathcal{J}_t^{\text{FS-EWC}}(\boldsymbol{\theta}; r) = \underbrace{\widehat{\mathcal{L}}_t^{\mathrm{CE}}(\boldsymbol{\theta})}_{\text{current-task fitting}} + \underbrace{\mathcal{R}_{\mathrm{F},t}^{\max}(\boldsymbol{\theta}; r)}_{\text{future-drift sensitivity shaping}} + \underbrace{\Omega_{t-1}(\boldsymbol{\theta})}_{\text{past-task protection}}. \tag{11}$$

Let $\boldsymbol{\epsilon}^\star \in \arg\max_{\|\boldsymbol{\epsilon}\|_2 \le r} \mathcal{R}_{\mathrm{F},t}(\boldsymbol{\theta} + \boldsymbol{\epsilon})$. At differentiable points, the gradient of Eq. 11 takes the form

$$\nabla_{\boldsymbol{\theta}} \mathcal{J}_t^{\text{FS-EWC}} = \boldsymbol{g}_t(\boldsymbol{\theta}) + \boldsymbol{h}_t(\boldsymbol{\theta} + \boldsymbol{\epsilon}^\star) + \lambda \boldsymbol{\Lambda}_{t-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}), \tag{12}$$

where

$$\boldsymbol{h}_t(\boldsymbol{u}) = \nabla_{\boldsymbol{u}} \mathcal{R}_{\mathrm{F},t}(\boldsymbol{u}) = \frac{2}{m_t} \sum_{i=1}^{m_t} \boldsymbol{H}_{t,i}(\boldsymbol{u}) \boldsymbol{g}_{t,i}(\boldsymbol{u}). \tag{13}$$

Thus, Fisher shaping introduces a Hessian–vector product correction that reduces neighborhood sensitivity of the current-task solution, while EWC retains historical parameter protection.

3.3.2 TRAJECTORY-BASED ANCHOR CONSTRUCTION

Sensitivity shaping determines how the current-task solution is learned, but continual learning also requires specifying which state is passed to subsequent tasks. Rather than retaining only the terminal iterate, we construct the task-t anchor by averaging selected learned checkpoints from its optimization trajectory. Let $\mathcal{K}_t$ denote the selected checkpoint indices, with $|\mathcal{K}_t| = K$. We define

$$\overline{\boldsymbol{\theta}}_t = \frac{1}{K} \sum_{k \in \mathbb{K}_t} \boldsymbol{\theta}_{t,k}, \qquad \boldsymbol{\theta}_t := \overline{\boldsymbol{\theta}}_t. \tag{14}$$

The average can be accumulated online and does not require storing the full optimization trajectory. To align the stored importance with the state actually inherited by future tasks, we estimate the current-task FIM at the retained anchor. Following Schwarz et al. (2018), we normalize the current-task FIM before consolidating it, and additionally apply a small isotropic damping term before consolidation: $\boldsymbol{\Lambda}_t = \boldsymbol{\Lambda}_{t-1} + \widehat{\boldsymbol{F}}_t$, where $\widehat{\boldsymbol{F}}_t$ denotes the normalized and damped current-task FIM. The pair $(\boldsymbol{\theta}_t, \boldsymbol{\Lambda}_t)$ initializes and regularizes the next task.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models and datasets We evaluate our method on vision and language continual learning benchmarks. Vision experiments use ViT-B/16 (Dosovitskiy, 2020) pretrained on ImageNet-21K. We evaluate on CIFAR-100 (Krizhevsky et al., 2009), ImageNet-R (Hendrycks et al., 2021a), ImageNet-A (Hendrycks et al., 2021b), and DomainNet (Peng et al., 2019), with ImageNet-A as the main benchmark. Additional evaluations cover five fine-grained datasets: FGVC Aircraft (Maji et al., 2013), Stanford Cars (Krause et al., 2013), CUB-200-2011 (Wah et al., 2011), Oxford Flowers (Nilsback & Zisserman, 2008), and Oxford-IIIT Pets (Parkhi et al., 2012), along with a *5Datasets* stream constructed from these datasets. Language experiments use Llama-3.2-1B and Llama-3.2-3B (Meta,

Table 1: Comparison results on ImageNet-R, ImageNet-A, DomainNet, and CIFAR-100. Results are reported in percentage with standard deviations. Bold and underline indicate the highest and second-highest scores.

Method	ImageNet-R (T=10)		ImageNet-A (T=10)		DomainNet (T=5)		CIFAR-100 (T=10)	
	FAA	AAA	FAA	AAA	FAA	AAA	FAA	AAA
Joint Train	$81.69_{(0.28)}$	$86.25_{(0.09)}$	$65.01_{(0.74)}$	$74.10_{(0.44)}$	$76.84_{(0.06)}$	$81.25_{(0.05)}$	$92.82_{(0.20)}$	$95.41_{(0.05)}$
Finetune	$60.42_{(1.64)}$	$73.18_{(0.32)}$	$16.31_{(7.89)}$	$30.44_{(13.18)}$	$51.46_{(0.47)}$	$67.08_{(1.13)}$	$69.49_{(0.50)}$	$80.35_{(0.87)}$
L2P	$71.26_{(0.44)}$	$76.13_{(0.46)}$	$42.94_{(1.27)}$	$51.40_{(1.95)}$	$70.26_{(0.25)}$	$75.83_{(0.98)}$	$83.18_{(1.20)}$	$87.69_{(1.05)}$
DualPrompt	$68.22_{(0.20)}$	$73.81_{(0.39)}$	$45.49_{(0.96)}$	$54.68_{(1.24)}$	$68.26_{(0.90)}$	$73.84_{(0.45)}$	$81.48_{(0.86)}$	$86.41_{(0.66)}$
EWC	$64.16_{(0.92)}$	$75.27_{(0.55)}$	$21.30_{(4.79)}$	$45.77_{(0.43)}$	$58.85_{(2.92)}$	$72.12_{(0.34)}$	$75.06_{(3.49)}$	$85.72_{(1.04)}$
FS-EWC	$68.46_{(1.37)}$	$76.60_{(2.45)}$	$26.89_{(0.98)}$	$46.08_{(2.37)}$	$64.77_{(0.53)}$	$74.87_{(0.59)}$	$77.32_{(0.59)}$	$85.70_{(0.27)}$
Vanilla LoRA	$64.87_{(0.73)}$	$75.57_{(0.23)}$	$40.01_{(1.32)}$	$58.28_{(0.85)}$	$69.79_{(0.11)}$	$77.44_{(0.08)}$	$82.99_{(0.84)}$	$89.74_{(0.58)}$
BiLoRA	$74.28_{(0.92)}$	$77.38_{(0.88)}$	$51.05_{(0.74)}$	$62.82_{(0.65)}$	$69.75_{(0.23)}$	$73.86_{(0.21)}$	$85.99_{(0.49)}$	$90.62_{(0.42)}$
InfLoRA	$\underline{74.41}_{(0.63)}$	$\underline{80.31}_{(0.60)}$	$50.75_{(1.33)}$	$64.36_{(1.01)}$	$71.01_{(0.05)}$	$77.75_{(0.03)}$	$86.34_{(0.76)}$	$91.33_{(0.48)}$
SD-LoRA	$72.93_{(2.76)}$	$79.80_{(0.15)}$	$55.23_{(0.94)}$	$66.10_{(0.54)}$	$71.27_{(0.14)}$	$77.70_{(0.07)}$	$86.77_{(0.30)}$	$90.96_{(0.30)}$
EWCLoRA	$72.58_{(0.58)}$	$78.38_{(0.69)}$	$\underline{56.42}_{(0.88)}$	$\underline{67.09}_{(0.62)}$	$\underline{73.38}_{(0.14)}$	$\underline{79.50}_{(0.13)}$	$\underline{87.91}_{(0.11)}$	$\mathbf{92.27}_{(0.19)}$
FS-EWCLoRA	$\mathbf{79.81}_{(0.21)}$	$\mathbf{83.51}_{(0.11)}$	$\mathbf{62.08}_{(1.62)}$	$\mathbf{71.66}_{(0.25)}$	$\mathbf{73.95}_{(0.04)}$	$\mathbf{79.50}_{(0.04)}$	$\mathbf{88.03}_{(0.10)}$	$\underline{92.00}_{(0.16)}$

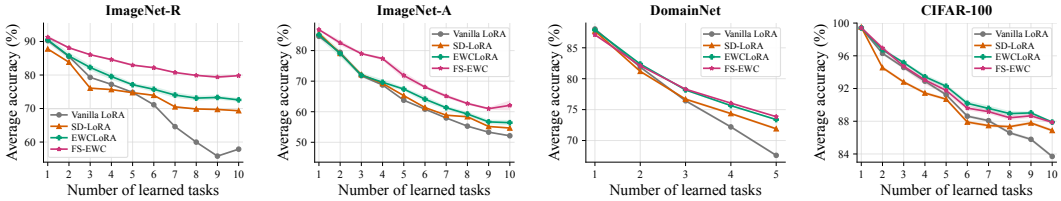


Figure 2: Stage-wise average accuracy A_t over all learned tasks under LoRA on four vision benchmarks.

2024) on a continual text-classification benchmark comprising AG News, Amazon Reviews, DBpedia, and Yahoo Answers (Zhang et al., 2015). Detailed training configurations are provided in Appendix B.1.

Metrics We report stage-wise average accuracy A_t over all tasks observed through learning stage t, together with final average accuracy (FAA), stage-averaged accuracy (AAA), and backward transfer (BWT). For the solution-shaping diagnostics, we additionally measure task-specific loss and accuracy, effective-weight displacement, empirical Fisher sensitivity, and displacement tolerance. We distinguish realized endpoint displacement d_t from controlled diagnostic displacement d, with $D_{1,t}^{\text{test}}$ and $\overline{D_1^+}(d)$ denoting the corresponding test-loss change and mean positive training-loss increase. Formal definitions and measurement protocols are provided in Appendix A.2.

4.2 Main Result

Overall Performance across Benchmarks and Model Configurations. We first evaluate whether the benefit of FS-EWC persists across training parameterizations, task domains, and model scales.

Table 1 reports results on the four standard vision benchmarks, with the corresponding LoRA-based performance trajectories shown in Fig. 2. FS-EWC improves FAA over the corresponding EWC baseline across all four benchmarks under both full-parameter fine-tuning and LoRA. The LoRA results on ImageNet-R and ImageNet-A are particularly notable. On ImageNet-R, EWCLoRA trails several competing LoRA methods, whereas FS-EWCLoRA improves FAA to 79.81%, surpassing all compared LoRA baselines. On ImageNet-A, EWCLoRA already leads the competing LoRA baselines, yet FS-EWCLoRA achieves a further gain, raising FAA from 56.42% to 62.08%. These results suggest that EWC parameter protection alone can be insufficient in challenging settings, whereas jointly shaping and protecting task solutions can yield substantial performance gains.

We next examine more heterogeneous visual recognition settings in Table 2. Across Aircraft, Cars, CUB-200, Flowers, and 5Datasets, FS-EWCLoRA improves FAA over its matched EWCLoRA baseline in every case. It achieves the highest FAA on Aircraft, Cars, and 5Datasets, and remains stronger on CUB-200 and Flowers. Thus, the improvement over EWC persists across substantially different class structures and task compositions.

Finally, we evaluate continual text classification with Llama-3.2-1B and Llama-3.2-3B across three task orders in Table 3. FS-EWCLoRA increases the mean FAA over EWCLoRA from 64.70% to

Table 2: Fine-grained and heterogeneous vision performance. FAA and stage-averaged AAA are reported in percentages. 5Datasets denotes the cross-dataset continual learning stream in the following order: CUB-200 → Aircraft → Cars → Flowers → Pets. The corresponding average-accuracy trajectories are provided in Fig. 6.

Method	FGVC-Aircraft		Stanford Cars		CUB-200		Oxford Flowers		5Datasets	
	FAA	AAA	FAA	AAA	FAA	AAA	FAA	AAA	FAA	AAA
Vanilla LoRA	39.85	38.42	31.74	41.67	58.31	69.61	52.92	66.49	66.90	68.85
InfLoRA	41.35	39.62	35.98	44.57	61.36	70.38	57.34	69.98	78.77	77.25
OLoRA	37.67	37.77	35.30	44.70	60.70	68.71	47.68	63.18	79.12	76.96
SDLoRA	25.96	26.63	34.89	43.28	62.01	69.52	49.05	64.59	74.42	74.10
CL-LoRA	29.35	26.54	34.01	41.86	**78.31**	**81.01**	**97.83**	**98.34**	67.05	61.97
EWCLoRA	40.64	38.44	34.56	43.94	61.34	70.36	52.02	65.36	73.30	72.93
FS-EWCLoRA	**46.66**	**43.41**	**53.04**	**56.26**	73.42	76.26	84.68	90.00	**83.24**	**81.19**

Table 3: Continual NLP benchmark performance of Llama-3.2-1B and Llama-3.2-3B across three task orders. FAA (%) is reported for each order, and Avg. denotes the average FAA across the three orders. Task-order are given in Table 9, and the corresponding trajectories are provided in Fig. 7.

Method	Llama-3.2-1B				Llama-3.2-3B			
	Order-1	Order-2	Order-3	Avg.	Order-1	Order-2	Order-3	Avg.
SeqLoRA	$65.55_{(0.59)}$	$58.53_{(0.86)}$	$67.83_{(1.65)}$	$63.97_{(0.96)}$	$67.21_{(0.92)}$	$67.00_{(0.37)}$	$69.59_{(3.10)}$	$67.93_{(0.60)}$
IncLoRA	$66.20_{(0.56)}$	$61.72_{(2.47)}$	$65.59_{(2.46)}$	$64.50_{(1.45)}$	$66.96_{(2.45)}$	$68.47_{(2.45)}$	$70.58_{(1.45)}$	$68.67_{(2.12)}$
OLoRA	$66.93_{(1.13)}$	$61.52_{(1.26)}$	$67.28_{(2.17)}$	$65.24_{(1.23)}$	$71.19_{(1.24)}$	$71.12_{(0.11)}$	$73.06_{(0.66)}$	$71.79_{(0.51)}$
TreeLoRA	$68.40_{(1.27)}$	$63.81_{(2.25)}$	$66.60_{(1.36)}$	$66.27_{(1.45)}$	$70.56_{(1.27)}$	$69.97_{(0.83)}$	$72.92_{(0.53)}$	$71.15_{(0.32)}$
EWCLoRA	$66.35_{(0.84)}$	$62.14_{(2.56)}$	$65.62_{(2.23)}$	$64.70_{(1.48)}$	$68.59_{(0.24)}$	$70.28_{(0.39)}$	$72.07_{(0.76)}$	$70.31_{(0.31)}$
FS-EWCLoRA	$\mathbf{69.57}_{(0.27)}$	$\mathbf{67.54}_{(1.04)}$	$\mathbf{70.35}_{(0.25)}$	$\mathbf{69.15}_{(0.49)}$	$\mathbf{72.57}_{(0.47)}$	$\mathbf{72.73}_{(0.67)}$	$\mathbf{73.93}_{(0.23)}$	$\mathbf{73.08}_{(0.40)}$

69.15% on Llama-3.2-1B and from 70.31% to 73.08% on Llama-3.2-3B, with improvements in five of the six model–order comparisons. These results indicate that the benefit is not restricted to vision or to a single model scale.

How Does Solution Shaping Affect the Stability–Plasticity Trade-off? We next examine whether the aggregate gains above reflect a more favorable stability–plasticity trade-off. Specifically, we ask whether improved stability is achieved merely by restricting future parameter movement, or by shaping the retained solution so that subsequent adaptation can proceed with less forgetting. To isolate the effect of Fisher shaping, each task only retains its terminal iterate without merging.

Fig. 3 first examines retention under actual subsequent learning. FS-EWCLoRA does not achieve improved stability by restricting parameter movement more strongly. At the final stage, its effective-weight displacement is larger than that of EWCLoRA (1.442 versus 1.254), while the Task 1 test-loss increase is substantially smaller (0.033 versus 0.229) and Task 1 accuracy is higher (86.64% versus 81.11%). The corresponding path Fisher is also much lower (0.015 versus 0.240). Thus, the retention gain is achieved despite greater realized movement, indicating that improved stability does not require stronger suppression of subsequent updates. This behavior is consistent with the finite-displacement analysis, which attributes old-task degradation jointly to displacement magnitude and the sensitivity encountered along that displacement.

To isolate the role of solution sensitivity, we next control the displacement magnitude directly. Fig. 4 perturbs the retained Task 1 solutions along matched random directions in the same effective LoRA weight. For the same perturbation radius, FS-EWCLoRA exhibits a smaller mean positive loss increase together with a lower empirical-Fisher trace throughout the measured range. Accordingly, for a fixed empirical retention budget, the shaped solution admits a larger measured displacement radius: for example, $\widehat{d}_\tau$ increases from 11 to 12.25 at $\tau = 0.01$, and is larger for ten of the eleven evaluated budgets. Hence, shaping enlarges the empirically retention-compatible region around the learned solution, providing more room for subsequent adaptation under the same retention tolerance.

Table 4 further separates solution shaping from historical protection. Comparing Vanilla LoRA with EWCLoRA shows the characteristic effect of EWC: the effective displacement d_2 decreases from 0.38 to 0.28, accompanied by a modest improvement in retained Task 1 accuracy , while Task 2 accuracy decreases slightly. Thus, historical protection primarily improves stability by restricting subsequent parameter movement, with a small cost to current-task adaptation.

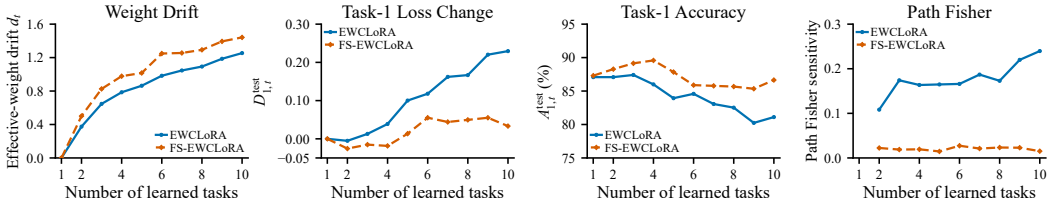


Figure 3: Less forgetting despite larger realized drift. Realized-drift diagnostics on ImageNet-A with LoRA. Relative to the Task 1 solution, we report effective-weight displacement d_t, Task 1 loss change $D^{\text{test}}_{1,t}$, Task 1 accuracy, and empirical-Fisher sensitivity averaged along the straight segment to each subsequent task endpoint.

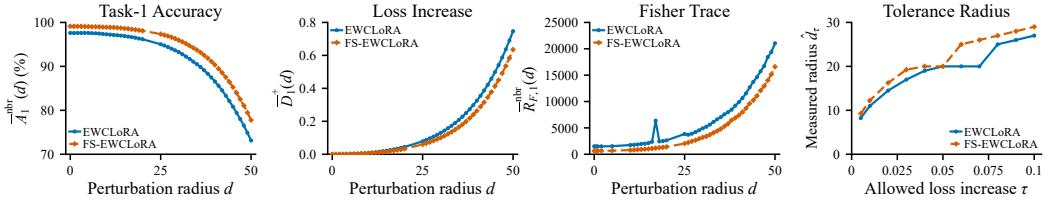


Figure 4: Lower neighborhood sensitivity and greater empirical displacement tolerance. Controlled-neighborhood diagnostics around the retained Task 1 solutions on ImageNet-A. Shared random directions are used to apply perturbations of radius d in the weight space. We report Task 1 training accuracy, mean positive training-loss increase, mean empirical-Fisher trace, and the measured tolerance radius under loss budget τ.

Table 4: Two-task diagnostics of solution shaping and historical protection with LoRA on ImageNet-A. C, F, and P denote current-task fitting, Fisher shaping, and historical EWC protection, respectively.

	Training objective		Task 1 endpoint		Task 2 endpoint				Task 2update
Variant	Task 1	Task 2	$a_{11}\uparrow$	$\mathrm{tr}\,\boldsymbol{F}_1^{\text{emp}}$	$a_{21}\uparrow$	$a_{22}\uparrow$	BWT↑	$A_2\uparrow$	d_2
Vanilla LoRA	C	C	$87.08_{(0.38)}$	$31.44_{(4.24)}$	$82.41_{(1.13)}$	$76.38_{(0.40)}$	$-4.67_{(1.36)}$	$79.39_{(0.74)}$	$0.38_{(0.02)}$
Fisher Shaping	C + F	C + F	$87.30_{(0.57)}$	$21.93_{(2.98)}$	$83.39_{(0.33)}$	$80.84_{(1.14)}$	$-3.91_{(0.65)}$	$82.12_{(0.59)}$	$0.50_{(0.01)}$
Task-1 only	C + F	C	$87.30_{(0.57)}$	$21.93_{(2.98)}$	$86.54_{(0.82)}$	$72.96_{(1.38)}$	$-0.76_{(1.36)}$	$79.75_{(1.10)}$	$0.36_{(0.01)}$
EWCLoRA	C	C + P	$87.08_{(0.38)}$	$31.44_{(4.24)}$	$82.95_{(0.76)}$	$76.25_{(0.60)}$	$-4.13_{(1.00)}$	$79.60_{(0.67)}$	$0.28_{(0.01)}$
FS-EWCLoRA	C + F	C + F + P	$87.30_{(0.57)}$	$21.93_{(2.98)}$	$83.50_{(0.50)}$	$79.92_{(1.04)}$	$-3.80_{(1.00)}$	$81.71_{(0.49)}$	$0.43_{(0.01)}$
Task-1 only	C + F	C + P	$87.30_{(0.57)}$	$21.93_{(2.98)}$	$86.32_{(0.86)}$	$73.36_{(1.20)}$	$-0.98_{(1.42)}$	$79.84_{(1.03)}$	$0.30_{(0.01)}$
Task-2 only	C	C + F + P	$87.08_{(0.38)}$	$31.44_{(4.24)}$	$81.32_{(0.50)}$	$81.89_{(1.04)}$	$-5.76_{(0.83)}$	$81.61_{(0.58)}$	$0.47_{(0.01)}$

However, Fisher shaping improves the retention–adaptation balance without relying on smaller parameter movement. Relative to Vanilla LoRA, Fisher Shaping improves both retained-task and current-task accuracy despite increasing d_2 from 0.38 to 0.50. The same pattern holds with EWC protection: FS-EWCLoRA improves a_{21} from 82.95% to 83.50% and a_{22} from 76.25% to 79.92%, while d_2 increases from 0.28 to 0.43. The stage-specific variants clarify why. Shaping Task 1 reduces the endpoint sensitivity proxy $tr\boldsymbol{F}_1^{\text{emp}}(\boldsymbol{\theta})$ from 31.44 to 21.93 and substantially improves its subsequent retention under either unprotected training ($82.41\% \to 86.54\%$) or the same C + P protection ($82.95\% \to 86.32\%$). Thus, shaping improves the solution passed forward for future retention. Conversely, shaping during Task 2 substantially improves current-task adaptation ($76.25\% \to 81.89\%$), although this shifts the balance away from immediate Task 1 retention. These results indicate that Fisher shaping serves two related roles: it forms solutions that are more suitable for subsequent retention and shapes the current-task solution during adaptation. Combined with historical protection, these effects yield a more favorable stability–plasticity balance than EWC alone.

Complementarity with FIM Representation We next examine whether the performance gains from solution shaping persist across different FIM schemes. We evaluate six combinations of three FIM computation schemes (batched empirical, per-example empirical, and model-sampled) and two representations (diagonal and Kronecker-factored).

Table 5 shows relatively modest performance differences across the evaluated FIM computations and representations. FS-EWC improves FAA over the corresponding EWC baseline by 6.08 to 9.07 percentage points across all six configurations. In these settings, the gains from FS-EWC exceed those obtained by refining the FIM computation or representation alone. Their consistency across

Table 5: Performance under different Fisher structures and computations. Under each structure, empirical FIM is computed using either batch gradients (Batch) or per-sample gradients (Per-sample), while Model denotes model-sampled FIM. Each entry reports FAA / AAA (%).

Method	Diagonal			Kronecker-factored		
	Batched	Per-example	Model-sampled	Batched	Per-example	Model-sampled
EWC	51.96 / 66.30	50.60 / 66.60	50.76 / 66.60	53.74 / 67.37	52.75 / 67.55	52.71 / 67.33
FS-EWC	59.29 / 70.69	59.67/ 70.88	59.62 / 70.86	59.82 / 71.00	59.95 / 71.16	59.62 / 71.12

Table 6: Constructing the anchor from learning trajectory. Anchor construction in FS-EWCLoRA on ImageNet-A, with current-task FIM estimated at the resulting anchor. Trajectory averaging achieves the highest mean AAA. Multi-seed results and checkpoint configurations are provided in Appendix B.5.

Retained state	FAA↑	AAA↑	BWT↑
Endpoint	59.23	70.63	−8.08
Late-stage	**59.43**	70.61	−7.68
Trajectory	59.34	**70.84**	−7.47
Dense trajectory	59.31	70.745	−7.44
Endpoint + initial	58.44	69.38	**−6.44**
Trajectory + initial	59.01	70.59	−7.33

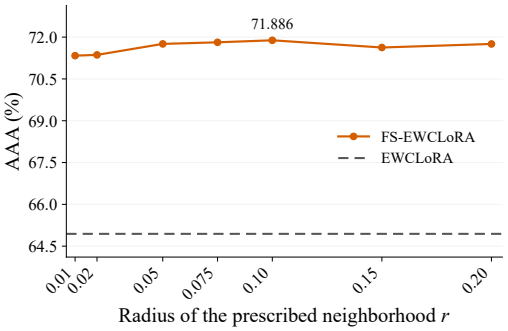


Figure 5: Performance across neighborhood radii. AAA of FS-EWCLoRA on ImageNet-A as the prescribed neighborhood radius r varies.

protection schemes highlights the complementary value of solution shaping: Fisher information can guide the formation of solutions for future retention as well as their subsequent protection. See Appendix B.4 for λ settings and sensitivity analyses.

4.3 Ablation and Hyperparameter Studies

Having isolated the roles of Fisher shaping and historical protection in the preceding mechanism study, we further examine two implementation choices of FS-EWC: retained-anchor construction and the prescribed neighborhood radius.

Retained-state construction. Table 6 shows that trajectory averaging achieves the highest AAA, while increasing checkpoint density provides no further improvement. Multi-checkpoint construction generally reduces forgetting relative to the endpoint. Including the initialization lowers FAA and AAA despite improving BWT. These results suggest that the choice of retained state matters beyond the training objective, but the benefit depends on which learned states are combined rather than simply on their number.

Neighborhood radius. Figure 5 evaluates the prescribed neighborhood radius. FS-EWCLoRA remains above the matched EWCLoRA baseline throughout the tested range, while AAA varies only moderately across different radii. These results indicate that the benefit of sensitivity shaping is not restricted to a narrowly tuned neighborhood radius.

5 Conclusion

EWC mitigates forgetting through Fisher-based parameter protection, but remains constrained by the stability–plasticity trade-off. Its historical protection is constructed from local parameter importance, whereas forgetting under subsequent learning depends on the sensitivity encountered after finite parameter displacement. Motivated by this gap, we introduced Fisher-Shaped EWC (FS-EWC), which incorporates Fisher information into current-task solution formation while preserving subsequent EWC protection. Our experiments show that shaped solutions can undergo larger subsequent displacement while incurring less forgetting, indicating a more favorable stability–plasticity balance. These findings suggest that future forgetting should be considered during learning.

AI USE STATEMENT

In this work, we used generative AI tools for assistance with mathematical formulation and proof development, methodological and experimental-design discussions, and interpretation of experimental results, and translation of manuscript content. We have not used generative AI tools to generate or modify experimental data, and tasks involving qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for manuscript organization, language editing, literature organization, reference formatting, and refinement of figures and captions.

We have reviewed all AI-assisted work. Mathematical statements and derivations were independently checked by the authors for correctness and consistency with the stated assumptions. Suggestions related to methodology, experimental design, and result interpretation were verified against the implemented methods, experimental protocols, and measured results. Literature-related content and references were checked against the original sources, and all AI-assisted manuscript text was manually reviewed and revised by the authors.

We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies algorithmic methods for replay-free continual learning and does not involve human subjects, personally identifiable information, or the collection of private data. All experiments are conducted on established public benchmarks. As with other machine learning methods, the behavior of the resulting models may inherit biases or limitations from the underlying datasets and pretrained models. Our work does not introduce additional data collection or deployment-specific risks beyond those associated with the datasets and model architectures considered.

REPRODUCIBILITY STATEMENT

We provide the information necessary to reproduce both the theoretical and empirical results of this work. The FS-EWC objective, neighborhood-sensitivity optimization, retained-anchor construction, and Fisher-based consolidation procedure are specified in the main methodology section. Complete derivations and proofs of the theoretical results are provided in the appendix. The appendix further documents dataset and task construction, model architectures, continual-learning protocols, baseline configurations, optimization hyperparameters, Fisher estimation procedures, evaluation metrics, and all settings used in the ablation and diagnostic studies. Code and experiment configurations will be released to facilitate independent reproduction of the reported results.

REFERENCES

Arslan Chaudhry, Puneet K. Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European Conference on Computer Vision*, pp. 532–547, 2018.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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, 2025. doi: 10.1109/CVPR52734.2025.02843.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8340–8349, October 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 (CVPR)*, pp. 15262–15271, June 2021b.

Ferenc Huszár. Note on the quadratic penalties in elastic weight consolidation. *Proceedings of the National Academy of Sciences*, 115(11):E2496–E2497, 2018. doi: 10.1073/pnas.1717042115.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://www.pnas.org/doi/abs/10.1073/pnas.1611835114.

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

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. ISBN 979-8-3503-5300-6. doi: 10.1109/CVPR52733.2024.02231. URL https://ieeexplore.ieee.org/document/10658274/.

Xialei Liu, Marc Masana, Luis Herranz, Joost van de Weijer, Antonio M. Lopez, and Andrew D. Bagdanov. Rotate your networks: Better weight consolidation and less catastrophic forgetting. In *2018 24th International Conference on Pattern Recognition*, pp. 2262–2268. IEEE, 2018. doi: 10.1109/ICPR.2018.8545895.

Yuhang Liu, Tao Li, Zhehao Huang, Zuopeng Yang, and Xiaolin Huang. Bi-lora: Efficient sharpness-aware minimization for fine-tuning large-scale models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 96782–96813, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/9c84feb75eae1ef6389f31b3ef050b6a-Paper-Conference.pdf.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft, 2013. URL https://arxiv.org/abs/1306.5151.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2408–2417. PMLR, 2015.

Meta. Llama 3.2 model card. Model card, 2024. URL https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 722–729. IEEE, 2008. doi: 10.1109/ICVGIP.2008.47.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505. IEEE, 2012. doi: 10.1109/CVPR.2012.6248092.

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

Yu-Yang Qian, Yuan-Ze Xu, Zhen-Yu Zhang, Peng Zhao, and Zhi-Hua Zhou. Treelora: Efficient continual learning via layer-wise loras guided by a hierarchical gradient-similarity tree, 2026. URL https://arxiv.org/abs/2506.10355.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4528–4537. PMLR, 2018.

Khadija Shaheen, Muhammad Abdullah Hanif, Osman Hasan, and Muhammad Shafique. Continual learning for real-world autonomous systems: Algorithms, challenges and frameworks. *Journal of Intelligent & Robotic Systems*, 105(1):9, 2022. ISSN 1573-0409. doi: 10.1007/s10846-022-01603-6. URL https://doi.org/10.1007/s10846-022-01603-6.

Gido M. van de Ven. On the computation of the fisher information in continual learning, 2025.

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.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383, 2024. doi: 10.1109/TPAMI.2024.3367329.

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 *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=L7ZBpZZ8Va.

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, 2025. URL https://arxiv.org/abs/2501.13198.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3987–3995. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/zenke17a.html.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.

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

A SUPPLEMENTARY THEORETICAL ANALYSIS

A.1 NOTATION

Table 7 collects the notation used below. All gradients, perturbations, and importance matrices use the same trainable parameterization.

Table 7: Key notation used in FS-EWC.

Symbol	Meaning
$\boldsymbol{\theta}$, $\boldsymbol{\theta}_t$	Candidate parameters and the retained task-t anchor.
$\boldsymbol{F}_t^{\mathrm{emp}}$, $\widehat{\boldsymbol{F}}_t$, ε_F	Empirical FIM and the normalized and damped current-task FIM, Fisher-normalization stabilizer.
$\boldsymbol{\Lambda}_t$, Ω_{t-1}	Consolidated importance and the inherited EWC penalty.
$\boldsymbol{\Delta}$, $\boldsymbol{\epsilon}$	Future parameter displacement and sensitivity-search perturbation.
$D_t(\boldsymbol{\theta}, \boldsymbol{\Delta})$	Task-loss change; positive values indicate deterioration.
$\mathcal{R}_{\mathrm{F},t}$, $\mathcal{R}_{\mathrm{F},t}^{\max}$	Fisher trace and its maximum over $\mathbb{U}_r(\boldsymbol{\theta})$.
$\boldsymbol{g}$, $\boldsymbol{H}$	Gradient and Hessian of the mean loss on a fixed batch $\mathbb{B}$.
$s(\boldsymbol{u})$, $\widehat{\boldsymbol{h}}(\boldsymbol{u})$	Probe step size and estimate of $2\boldsymbol{H}(\boldsymbol{u})\boldsymbol{g}(\boldsymbol{u})$.
$\boldsymbol{z}$, $\boldsymbol{d}_{t,\mathbb{B}}$	Selected neighborhood point and numerical update direction.
$\mathbb{K}_t$, K, $\overline{\boldsymbol{\theta}}_t$	Selected checkpoint indices, their number, and their parameter average.
β, γ, λ	Damping, decay coefficient, and historical-regularization.
r,ρ, ξ	Neighborhood radius, finite-difference probe scale. and normalization stabilizer.

A.2 EVALUATION QUANTITIES.

Let $a_{t,i}$ denote the test accuracy on Task i after learning through Task t, where $1 \le i \le t \le T$. The average accuracy after learning t tasks is

$$A_t = \frac{1}{t}\sum_{i=1}^{t} a_{t,i}. \tag{15}$$

This is the stage-wise average accuracy reported in the learning curves. We summarize continual-learning performance using final average accuracy (FAA), average accuracy across learning stages (AAA), and backward transfer (BWT):

$$\mathrm{FAA} = A_T, \quad \mathrm{AAA} = \frac{1}{T}\sum_{t=1}^{T} A_t \quad \mathrm{BWT} = \frac{1}{T-1}\sum_{i=1}^{T-1}\left(a_{T,i} - a_{i,i}\right). \tag{16}$$

A more negative BWT indicates greater forgetting.

For the solution-shaping diagnostics, we additionally characterize parameter displacement and solution sensitivity in the effective LoRA update space. Specifically, we write

$$\mathbf{w}_t = \mathrm{vec}\left(\{B_{\ell c,t}A_{\ell c,t}\}_{\ell,\ c\in\{Q,V\}}\right) \tag{17}$$

for the vectorized effective LoRA weights retained after Task t. For realized subsequent learning, define

$$\Delta\mathbf{w}_t = \mathbf{w}_t - \mathbf{w}_1, \qquad d_t = \|\Delta\mathbf{w}_t\|_2, \tag{18}$$

and the signed Task 1 test-loss change

$$D_{1,t}^{\mathrm{test}} = \widehat{\mathcal{L}}_{1,\mathrm{test}}(\mathbf{w}_t) - \widehat{\mathcal{L}}_{1,\mathrm{test}}(\mathbf{w}_1). \tag{19}$$

For $d_t > 0$, let $\mathbf{v}_t = \Delta\mathbf{w}_t/d_t$. The displacement-normalized directional Fisher sensitivity along the straight segment connecting the two endpoints is

$$\overline{\mathcal{S}}_{\mathrm{F},1}^{\mathrm{path}}(t) = \int_0^1 \mathbf{v}_t^\top \boldsymbol{F}_{1,\mathrm{tr}}^{\mathrm{emp}}(\mathbf{w}_1 + s\Delta\mathbf{w}_t)\mathbf{v}_t\, ds. \tag{20}$$

For the controlled-neighborhood diagnostic, let d denote the prescribed perturbation radius and $\mathbf{u}_k$ a unit perturbation direction. The Task 1 training-loss change along direction k is

$$D_{1,k}^{\mathrm{nbr}}(d) = \widehat{\mathcal{L}}_{1,\mathrm{tr}}(\mathbf{w}_1 + d\mathbf{u}_k) - \widehat{\mathcal{L}}_{1,\mathrm{tr}}(\mathbf{w}_1). \tag{21}$$

We report

$$\overline{D}_1^{+}(d) = \frac{1}{K_D}\sum_{k=1}^{K_D}[D_{1,k}^{\mathrm{nbr}}(d)]_+, \qquad \overline{\mathcal{R}}_{\mathrm{F},1}^{\mathrm{nbr}}(d) = \frac{1}{K_F}\sum_{k=1}^{K_F}\operatorname{tr}\boldsymbol{F}_{1,\mathrm{tr}}^{\mathrm{emp}}(\mathbf{w}_1 + d\mathbf{u}_k). \tag{22}$$

For an allowed mean positive loss increase τ, the measured tolerance radius over the evaluated grid $0 = d_0 < d_1 < \cdots < d_J$ is

$$\widehat{d}_\tau = \max\left\{d_j : \overline{D}_1^{+}(d_i) \le \tau \text{ for all } i \le j\right\}. \tag{23}$$

Because $\widehat{d}_\tau$ is evaluated only on the measured grid, it is an empirical summary rather than a certificate over the continuous neighborhood.

For the two-task diagnostic, A_2 is the corresponding stage-average accuracy from Eq. equation 15, i.e., $A_2 = (a_{21} + a_{22})/2$. In this case, $\mathrm{BWT} = a_{21} - a_{11}$ and $d_2 = \|\mathbf{w}_2 - \mathbf{w}_1\|_2$.

A.3 Proof of Finite-Displacement Sensitivity

Fix $\|\boldsymbol{\Delta}\|_2 \le r$. The segment $\boldsymbol{\theta} + \tau\boldsymbol{\Delta}$, $\tau \in [0,1]$, lies in $\mathbb{U}_r(\boldsymbol{\theta})$, so the differentiability assumption in Proposition 1 gives

$$D_t(\boldsymbol{\theta}, \boldsymbol{\Delta}) = \int_0^1 \frac{1}{m_t}\sum_{i=1}^{m_t}\boldsymbol{g}_{t,i}(\boldsymbol{\theta} + \tau\boldsymbol{\Delta})^\top\boldsymbol{\Delta}\,\mathrm{d}\tau. \tag{24}$$

Applying Cauchy–Schwarz to the integral and then to the sample average yields

$$\begin{aligned} |D_t(\boldsymbol{\theta}, \boldsymbol{\Delta})|^2 &\le \int_0^1 \left[\frac{1}{m_t}\sum_{i=1}^{m_t}\boldsymbol{g}_{t,i}(\boldsymbol{\theta} + \tau\boldsymbol{\Delta})^\top\boldsymbol{\Delta}\right]^2 \mathrm{d}\tau \\ &\le \int_0^1 \frac{1}{m_t}\sum_{i=1}^{m_t}\left[\boldsymbol{g}_{t,i}(\boldsymbol{\theta} + \tau\boldsymbol{\Delta})^\top\boldsymbol{\Delta}\right]^2 \mathrm{d}\tau \\ &= \int_0^1 \boldsymbol{\Delta}^\top\boldsymbol{F}_t^{\mathrm{emp}}(\boldsymbol{\theta} + \tau\boldsymbol{\Delta})\boldsymbol{\Delta}\,\mathrm{d}\tau. \end{aligned} \tag{25}$$

This proves the proposition.

A.4 FS-EWC from a Bayesian Perspective

The Bayesian perspective provides a complementary interpretation of both components of FS-EWC: accounting for a parameter distribution when learning its center, and selecting the center retained for subsequent tasks.

EWC from a Bayesian perspective. Sequential Bayesian learning maintains a distribution over model parameters. For conditionally independent task data,

$$p(\boldsymbol{\theta} \mid \mathbb{S}_{1:t}) \propto p(\boldsymbol{\theta} \mid \mathbb{S}_{1:t-1})p(\mathbb{S}_t \mid \boldsymbol{\theta}). \tag{26}$$

Thus, the posterior after task $t-1$ serves as the prior for learning task t, carrying historical information without requiring the original data.

The exact posterior is generally intractable for neural networks. EWC approximates the inherited parameter distribution by a local Gaussian surrogate (Kirkpatrick et al., 2017; Huszár, 2018):

$$q_{t-1}(\boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{\theta}; \boldsymbol{\theta}_{t-1}, \boldsymbol{\Lambda}_{t-1}^{-1}\right). \tag{27}$$

Using this surrogate as the prior, a MAP-style current-task update is

$$\boldsymbol{\theta}_t^{\mathrm{MAP}} = \arg\min_{\boldsymbol{\theta}} \left[\widehat{\mathcal{L}}_t^{\mathrm{CE}}(\boldsymbol{\theta}) + \frac{\lambda}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1})^\top \boldsymbol{\Lambda}_{t-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}) \right] = \mathcal{J}_t^{\mathrm{EWC}}(\boldsymbol{\theta}), \tag{28}$$

where λ absorbs the practical scaling between the mean current-task loss and the inherited precision. This recovers the EWC objective in equation 2. Hence, EWC retains a distribution-derived representation of historical uncertainty, but performs current-task learning through a point estimate of the new parameters.

Optimizing the distribution. The underlying Bayesian update is distributional rather than pointwise. A corresponding variational objective is

$$\mathcal{V}_t(Q) = \mathbb{E}_{\boldsymbol{u}\sim Q}[\widehat{\mathcal{L}}_t^{\mathrm{CE}}(\boldsymbol{u})] + \lambda \mathrm{KL}(Q \| q_{t-1}), \tag{29}$$

where Q is a candidate current-task parameter distribution. The first term evaluates current-task performance across parameter values represented by Q, whereas standard EWC evaluates the current-task loss only at its MAP candidate.

To expose this difference, consider a fixed-shape local family $Q_{\boldsymbol{\theta}}$ represented by

$$\boldsymbol{u} = \boldsymbol{\theta} + \boldsymbol{\epsilon}, \qquad \mathbb{E}[\boldsymbol{\epsilon}] = \mathbf{0}, \qquad \|\boldsymbol{\epsilon}\|_2 \le r. \tag{30}$$

Its expected current-task loss decomposes as

$$\mathbb{E}_{\boldsymbol{u}\sim Q_{\boldsymbol{\theta}}}[\widehat{\mathcal{L}}_t^{\mathrm{CE}}(\boldsymbol{u})] = \widehat{\mathcal{L}}_t^{\mathrm{CE}}(\boldsymbol{\theta}) + \mathbb{E}[D_t(\boldsymbol{\theta}, \boldsymbol{\epsilon})]. \tag{31}$$

The second term measures the loss variation over parameter values around the candidate center and is not controlled by center loss alone.

For this fixed-shape, zero-mean family, the historical KL term has the same center-dependent part as EWC:

$$\lambda \mathrm{KL}(Q_{\boldsymbol{\theta}} \| q_{t-1}) = \Omega_{t-1}(\boldsymbol{\theta}) + C, \tag{32}$$

where C does not depend on $\boldsymbol{\theta}$. Using the finite-displacement Fisher bound,

$$\mathbb{E}[D_t(\boldsymbol{\theta}, \boldsymbol{\epsilon})] \le r\sqrt{\mathcal{R}_{\mathrm{F},t}^{\max}(\boldsymbol{\theta}; r)} \le \mathcal{R}_{\mathrm{F},t}^{\max}(\boldsymbol{\theta}; r) + \frac{r^2}{4}. \tag{33}$$

Therefore,

$$\mathcal{V}_t(Q_{\boldsymbol{\theta}}) \le \mathcal{J}_t^{\mathrm{FS-EWC}}(\boldsymbol{\theta}; r) + C + \frac{r^2}{4}. \tag{34}$$

For fixed distribution shape, r, the parameter-dependent upper surrogate is exactly the FS-EWC objective. Thus, FS-EWC can be viewed as extending EWC's pointwise MAP update toward distribution-aware center selection: the center is chosen not only for its own task loss and historical compatibility, but also for controlled loss variation over its surrounding parameter distribution.

Selecting the retained center. At task completion, let Q_t^{traj} assign equal mass to the selected late-stage checkpoints $\boldsymbol{\theta}_{t,k}$, $k \in \mathbb{K}_t$. For any fixed positive definite matrix $\boldsymbol{A}$, fitting a Gaussian center to this empirical distribution yields

$$\arg\min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{u}\sim Q_t^{\mathrm{traj}}}[-\log \mathcal{N}(\boldsymbol{u}; \boldsymbol{\theta}, \boldsymbol{A}^{-1})] \quad = \frac{1}{K}\sum_{k\in\mathbb{K}_t} \boldsymbol{\theta}_{t,k} = \overline{\boldsymbol{\theta}}_t. \tag{35}$$

The equality follows by minimizing the associated quadratic in the center and recovers the retained-anchor rule $\boldsymbol{\theta}_t := \overline{\boldsymbol{\theta}}_t$. It treats the selected checkpoints as an empirical distribution, without assuming they are exact posterior samples. Fisher importance is then estimated at the retained anchor and consolidated as specified in Section 3.3.2. Thus, the distributional view motivates both learning the center through Fisher shaping and selecting the retained center through trajectory averaging.

A.5 Practical Implementation

A.5.1 Empirical Fisher Computation

Several Fisher variants are used in continual learning, differing in the distribution over labels and in how expectations are approximated. Our theoretical development uses the per-example empirical Fisher in Eq. equation 3. We focus here on the specific approximation used by our method and refer to Martens (2020); van de Ven (2025) for broader treatments of Fisher definitions and computation.

The theoretical sensitivity target is defined using the per-example empirical-Fisher trace, whereas practical mini-batch optimization uses the squared gradient of the batch-averaged loss. For a fixed batch $\mathbb{B}$ with $b = |\mathbb{B}|$, let

$$\boldsymbol{g}(\boldsymbol{u}) = \frac{1}{b}\sum_{i\in\mathbb{B}} \boldsymbol{g}_{t,i}(\boldsymbol{u}).$$

The two quantities satisfy the exact decomposition

$$\frac{1}{b}\sum_{i\in\mathbb{B}} \|\boldsymbol{g}_{t,i}(\boldsymbol{u})\|_2^2 = \|\boldsymbol{g}(\boldsymbol{u})\|_2^2 + \frac{1}{b}\sum_{i\in\mathbb{B}} \|\boldsymbol{g}_{t,i}(\boldsymbol{u}) - \boldsymbol{g}(\boldsymbol{u})\|_2^2. \tag{36}$$

Thus, the squared batch gradient omits the nonnegative dispersion of per-example gradients and is generally not identical to the empirical-Fisher trace. We use it as a computational surrogate because the batch-averaged gradient is directly available during stochastic optimization.

A.5.2 Fisher Post-processing and Consolidation

Following prior work (Schwarz et al., 2018) on Fisher normalization, we normalize the current-task Fisher importance to reduce arbitrary task-dependent differences in its absolute scale and make consolidation depend primarily on the relative importance of protected parameters. Specifically, we use global trace normalization over all protected parameter coordinates:

$$\widetilde{\boldsymbol{F}}_t = \mathcal{N}(\mathrm{Diag}(\boldsymbol{F}_t^{\mathrm{emp}}(\boldsymbol{\theta}_t))) = \frac{\mathrm{Diag}(\boldsymbol{F}_t^{\mathrm{emp}}(\boldsymbol{\theta}_t))}{\mathrm{tr}(\mathrm{Diag}(\boldsymbol{F}_t^{\mathrm{emp}}(\boldsymbol{\theta}_t))) + \xi_F}, \qquad \xi_F = 10^{-12}. \tag{37}$$

The trace is computed jointly over all protected coordinates rather than independently for individual layers. For LoRA adaptation, these coordinates correspond to the effective weight updates ΔW. Because all Fisher entries share the same normalization factor, their relative importance is preserved, while for nondegenerate Fisher statistics, $\mathrm{tr}(\widetilde{\boldsymbol{F}}_t) \approx 1$.

We then apply isotropic damping,

$$\widehat{\boldsymbol{F}}_t = (1-\beta)\widetilde{\boldsymbol{F}}_t + \beta\frac{\boldsymbol{I}}{d}, \qquad \beta \in [0,1], \tag{38}$$

where d is the total number of protected coordinates and $\beta = 0.2$ by default. Since $\mathrm{tr}(\boldsymbol{I}/d) = 1$, damping preserves the normalized trace, i.e., $\mathrm{tr}(\widehat{\boldsymbol{F}}_t) \approx 1$.

Following recursive consolidation as in online EWC (Schwarz et al., 2018), we update

$$\boldsymbol{\Lambda}_t = \gamma\boldsymbol{\Lambda}_{t-1} + \widehat{\boldsymbol{F}}_t, \qquad \gamma \in [0,1], \tag{39}$$

with $\gamma = 1$ by default.

A.5.3 Practical Sensitivity Solver

For a fixed task t and training batch $\mathbb{B}$, let

$$\boldsymbol{g}(\boldsymbol{u}) = \nabla_{\boldsymbol{u}}\widehat{\mathcal{L}}_{t,\mathbb{B}}^{\mathrm{CE}}(\boldsymbol{u}), \qquad \boldsymbol{H}(\boldsymbol{u}) = \nabla_{\boldsymbol{u}}^2\widehat{\mathcal{L}}_{t,\mathbb{B}}^{\mathrm{CE}}(\boldsymbol{u}).$$

For twice continuously differentiable batch losses, the derivative of the squared-gradient is

$$\nabla_{\boldsymbol{u}}\|\boldsymbol{g}(\boldsymbol{u})\|_2^2 = 2\boldsymbol{H}(\boldsymbol{u})\boldsymbol{g}(\boldsymbol{u}). \tag{40}$$

Sensitivity-gradient estimator.

Sensitivity-gradient estimator. We approximate this derivative using a normalized finite difference probe with a fixed unit scale. At a generic parameter point $\boldsymbol{u}$, define

$$\begin{aligned} \mathbf{s}(\boldsymbol{u}) &= \frac{\boldsymbol{g}(\boldsymbol{u})}{\|\boldsymbol{g}(\boldsymbol{u})\|_2 + \xi}, \\ \widehat{\boldsymbol{h}}(\boldsymbol{u}) &= 2\big(\|\boldsymbol{g}(\boldsymbol{u})\|_2 + \xi\big)\,\big[\boldsymbol{g}\big(\boldsymbol{u} + \rho\mathbf{s}(\boldsymbol{u})\big) - \boldsymbol{g}(\boldsymbol{u})\big]\,/\rho. \end{aligned} \tag{41}$$

Here, $\mathbf{s}(\boldsymbol{u})$ specifies the normalized probe direction, ρ controls the finite-difference probe scale, and $\xi > 0$ stabilizes the normalization. Since $\|\rho\mathbf{s}(\boldsymbol{u})\|_2 \le \rho$, ρ determines the maximum probe displacement. Unless otherwise specified, we set $\rho = 1$. When the Hessian varies slowly along the probe segment,

$$\widehat{\boldsymbol{h}}(\boldsymbol{u}) \approx 2\boldsymbol{H}(\boldsymbol{u})\boldsymbol{g}(\boldsymbol{u}) = \nabla_{\boldsymbol{u}}\|\boldsymbol{g}(\boldsymbol{u})\|_2^2. \tag{42}$$

Single-step neighborhood search and update. We approximate the inner maximization by a single normalized ascent step. The estimated sensitivity gradient determines the search direction, while the prescribed radius r controls the extent of the neighborhood search:

$$\begin{aligned} \widehat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}) &= r\frac{\widehat{\boldsymbol{h}}(\boldsymbol{\theta})}{\|\widehat{\boldsymbol{h}}(\boldsymbol{\theta})\|_2 + \xi}, \\ \boldsymbol{z} &= \boldsymbol{\theta} + \widehat{\boldsymbol{\epsilon}}(\boldsymbol{\theta}), \\ \boldsymbol{d}_{t,\mathbb{B}} &= \boldsymbol{g}(\boldsymbol{\theta}) + \widehat{\boldsymbol{h}}(\boldsymbol{z}) + \lambda\boldsymbol{\Lambda}_{t-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_{t-1}). \end{aligned} \tag{43}$$

By construction, $\|\widehat{\epsilon}(\boldsymbol{\theta})\|_2 \le r$, so the selected point $\boldsymbol{z}$ remains within the prescribed neighborhood of $\boldsymbol{\theta}$. The first estimator call provides a first-order approximation to the inner search direction used to select this neighborhood point, while the second approximates the sensitivity gradient at $\boldsymbol{z}$ and supplies the outer shaping correction.

A.5.4 Complete Learning Procedure

Algorithm 1 summarizes the complete FS-EWC procedure, including within-task sensitivity shaping, retained-anchor construction, and task-boundary Fisher consolidation.

Algorithm 1 Fisher-Shaped EWC (FS-EWC)

Require: Initial parameters $\boldsymbol{\theta}_0$; datasets $\mathbb{S}_1, \ldots, \mathbb{S}_T$; $\lambda, \beta, r, \rho, \gamma$; checkpoint sets $\mathbb{K}_t$; base optimizer.

Initialize $\boldsymbol{\Lambda}_0 \leftarrow \mathbf{0}$.
for $t = 1, \ldots, T$ **do**
  Set $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_{t-1}$ and reset the running checkpoint average.
  for each training batch $\mathbb{B}$ from $\mathbb{S}_t$ **do**
    Evaluate $\widehat{\boldsymbol{h}}(\boldsymbol{\theta})$ using Eq. 41.
    Construct the neighborhood point $\boldsymbol{z}$ using Eq. 43.
    Evaluate $\widehat{\boldsymbol{h}}(\boldsymbol{z})$ using the same batch.
    Form $\boldsymbol{d}_{t,\mathbb{B}}$ using Eq. 43.
    Update the unperturbed parameters $\boldsymbol{\theta}$ with the base optimizer using $\boldsymbol{d}_{t,\mathbb{B}}$.
    Accumulate $\boldsymbol{\theta}$ when the current checkpoint belongs to $\mathbb{K}_t$.
  end for
  Construct the retained anchor $\boldsymbol{\theta}_t$ using Eq. 14.
  Estimate $\boldsymbol{F}_t^{\mathrm{emp}}(\boldsymbol{\theta}_t)$ using the current-task data.
  Form the normalized and damped current-task importance $\widehat{\boldsymbol{F}}_t$ using Eqs. 37 and 38.
  Consolidate historical importance: $\boldsymbol{\Lambda}_t \leftarrow \gamma\boldsymbol{\Lambda}_{t-1} + \widehat{\boldsymbol{F}}_t$ using Eq. 39.
end for

Ensure: Retained parameters $\boldsymbol{\theta}_T$ and consolidated importance $\boldsymbol{\Lambda}_T$.

A.5.5 Approximation Error Analysis

Finite-Difference Approximation The estimator in Eq. 41 approximates the exact squared-gradient derivative at a fixed parameter point. By the fundamental theorem of calculus,

$$\widehat{\boldsymbol{h}}(\boldsymbol{u}) = 2\int_0^1 \boldsymbol{H}\big(\boldsymbol{u} + \tau\rho\mathbf{s}(\boldsymbol{u})\big)\boldsymbol{g}(\boldsymbol{u})\,\mathrm{d}\tau. \tag{44}$$

Hence, continuity of the Hessian implies $\widehat{\boldsymbol{h}}(\boldsymbol{u}) \to 2\boldsymbol{H}(\boldsymbol{u})\boldsymbol{g}(\boldsymbol{u}) \quad \text{as } \rho \to 0.$

If the Hessian is L_H-Lipschitz along the probe segment, then

$$\left\|\widehat{\boldsymbol{h}}(\boldsymbol{u}) - 2\boldsymbol{H}(\boldsymbol{u})\boldsymbol{g}(\boldsymbol{u})\right\|_2 \le L_H|\rho|\frac{\|\boldsymbol{g}(\boldsymbol{u})\|_2^2}{\|\boldsymbol{g}(\boldsymbol{u})\|_2 + \xi}. \tag{45}$$

Thus, the finite-difference estimator has first-order truncation error with respect to the probe scale ρ.

Single-Step Inner-Search Approximation

The neighborhood point $\boldsymbol{z}$ is obtained from a single first-order ascent step rather than by solving the inner maximization exactly. For a perturbation $\boldsymbol{\epsilon}$ around $\boldsymbol{\theta}$,

$$\|\boldsymbol{g}(\boldsymbol{\theta}+\boldsymbol{\epsilon})\|_2^2 = \|\boldsymbol{g}(\boldsymbol{\theta})\|_2^2 + \left[2\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})\right]^\top \boldsymbol{\epsilon} + o(\|\boldsymbol{\epsilon}\|_2). \tag{46}$$

Therefore, the normalized direction of $2\boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{g}(\boldsymbol{\theta})$ maximizes the linearized sensitivity objective over the radius-r ball. The practical solver replaces this exact sensitivity gradient with its finite-difference estimate and uses Eq. 43.

Let $\boldsymbol{\epsilon}^\star \in \arg\max_{\|\boldsymbol{\epsilon}\|_2 \le r} \|\boldsymbol{g}(\boldsymbol{\theta}+\boldsymbol{\epsilon})\|_2^2$ denote an exact inner maximizer. The practical point $\boldsymbol{z}$ need not coincide with $\boldsymbol{\theta} + \boldsymbol{\epsilon}^\star$ because it is obtained from a single local ascent step. Consequently, the practical solver introduces two distinct approximations: finite-difference estimation of the sensitivity gradient and single-step approximation of the neighborhood maximization. The probe scale ρ affects the former, whereas the neighborhood radius r controls the search region of the latter.

B Supplementary Experiments

B.1 Experimental Configurations

We organize the experimental setup into three groups: (i) standard vision benchmarks, (ii) fine-grained and heterogeneous vision benchmarks, and (iii) continual text classification. For published baselines, we follow the configurations reported by the original papers or released implementations whenever available. For EWC and FS-EWC, we report the configurations associated with the corresponding experimental runs.

Standard Vision LoRA Benchmarks We evaluate on CIFAR-100, DomainNet, ImageNet-R, and ImageNet-A following the EWC-LoRA (Zheng et al., 2026). CIFAR-100, ImageNet-R, and ImageNet-A each contain 10 sequential tasks, whereas DomainNet contains 5 tasks.

We use the same ViT-B/16 backbone, pretrained initialization, LoRA placement, and dataset-specific ranks as the reference setting, with ranks 10, 30, 10, and 10, respectively. The reported Vanilla LoRA, InfLoRA (Liang & Li, 2024), SD-LoRA (Wu et al., 2025), CL-LoRA (He et al., 2025), and BiLoRA (Liu et al., 2026) results are taken directly from Zheng et al. (2026). We refer readers to that work for their optimizer, learning-rate, classifier, and method-specific regularization settings.

FS-EWCLoRA follows the same task construction, backbone, LoRA placement, and dataset-specific ranks as the above benchmark. We use SGD with batch size 128 and cosine learning-rate scheduling. The shaping radius is $r = 0.05$ across datasets. The probe radius is $\rho = 0.1$ on CIFAR-100, and $\rho = 1$ on ImageNet-A, ImageNet-R and DomainNet. The damping coefficient is 0.2. Checkpoints are collected every 5, 5, 20, and 20 epochs, for CIFAR-100, DomainNet, ImageNet-R, and ImageNet-A, respectively.

We additionally evaluate EWC and FS-EWC under full-model fine-tuning on the same four benchmarks. All runs use ViT-B/16, SGD, learning rate 0.01, batch size 128, cosine scheduling, zero

momentum, and zero weight decay. The numbers of epochs per task are 20, 5, 30, and 20 for CIFAR-100, DomainNet, ImageNet-R, and ImageNet-A, respectively. The EWC and FS-EWC regularization coefficients are 4×10^4, 1×10^5, 8×10^4, and 2×10^5 for CIFAR-100, DomainNet, ImageNet-R, and ImageNet-A, respectively. For FS-EWC, the shaping radius r / finite-difference probe scale ρ are set to $0.005/0.10$, $0.05/0.20$, $0.05/0.20$, and $0.01/0.20$, with retained-checkpoint intervals of 5, 1, 10, and 20 epochs, respectively. All full-model FS-EWC runs use damping coefficient 0.2. Fisher statistics are estimated from at most 100 current-task batches and exclude the classification head.

Fine-Grained and Heterogeneous Vision Benchmarks We further evaluate LoRA-based continual learning on Aircraft, Cars, CUB-200, and Flowers, together with the heterogeneous 5Datasets benchmark. Aircraft, Cars, CUB-200, and Flowers are divided into 10 sequential tasks, while Pets contains 9 tasks. For 5Datasets, we follow the task order CUB-200 $\rightarrow$ Aircraft $\rightarrow$ Cars $\rightarrow$ Flowers $\rightarrow$ Pets. The detailed task partitions are summarized in Table 8.

Table 8: Task partitions for the fine-grained and heterogeneous vision benchmarks.

Dataset	Classes	Tasks	Task partition
Aircraft	100	10	10×10
Cars	196	10	$16+9\times20$
CUB-200	200	10	20×10
Flowers	102	10	$12+9\times10$
5Datasets	635	5	$200/100/196/102/37$

All methods use ViT-B/16 initialized from an ImageNet-1K fine-tuned checkpoint. LoRA is applied to the query and value projections with rank 16. We use SGD with batch size 128, zero momentum, zero weight decay, and cosine learning-rate scheduling. For the individual fine-grained datasets, models are trained for 40 epochs per task. The learning rates are 0.05 for Aircraft and Cars, 0.01 for CUB-200, and 0.0025 for Flowers. For 5Datasets, we use learning rate 0.02.

For both EWC-LoRA and FS-EWC-LoRA, we use the same LoRA rank, optimizer, learning rate, and training schedule described above, with the EWC regularization coefficient fixed to $\lambda = 2000$ on all benchmarks. FS-EWCLoRA additionally uses shaping radius $r = 0.05$ and finite-difference probe scale $\rho = 0.2$. The current runs retain only the task-end model, so no multi-checkpoint averaging is applied in this experiment.

Continual Text Classification We evaluate on four text-classification datasets from the continual learning benchmark of Zhang et al. (2015): AG News, Amazon Reviews, DBpedia, and Yahoo Answers. We consider three task orders to evaluate robustness to different task sequences as shown in Table 9.

Table 9: Task orders for continual text classification.

Order	Task sequence
1	DBpedia $\rightarrow$ Amazon $\rightarrow$ Yahoo $\rightarrow$ AG News
2	DBpedia $\rightarrow$ Amazon $\rightarrow$ AG News $\rightarrow$ Yahoo
3	Yahoo $\rightarrow$ Amazon $\rightarrow$ AG News $\rightarrow$ DBpedia

We use Llama-3.2-1B and Llama-3.2-3B with LoRA applied to the query and value projections. All experiments use LoRA rank 8, scaling factor 32, dropout 0.1, BF16 precision, and one training epoch per task. The first task uses learning rate 10^{-3} and subsequent tasks use 10^{-4} with a constant schedule and no warmup. The maximum input length is 512, the target/generation length is 50, gradient clipping is 1.0, and the effective batch size is 64 following OLoRA (Wang et al., 2023). AdamW is used for optimization.

For EWC-LoRA, the EWC coefficient is $\lambda = 0.5$ for Llama-3.2-1B and $\lambda = 1$ for Llama-3.2-3B. FS-EWC-LoRA uses $\lambda = 1$ for both model sizes, together with shaping radius $r = 0.02$ and finite-difference probe scale $\delta = 0.2$. The remaining shaping coefficients are fixed across all task orders. Following EWCLoRA, we also adopt TreeLoRA (Qian et al., 2026) for comparison.

B.2 Additional Experimental Result

Stage-wise performance on fine-grained and heterogeneous benchmarks. Figure 6 complements the aggregate results in Table 2 by showing how average accuracy evolves as new tasks are introduced. FS-EWCLoRA exhibits particularly strong late-stage performance on Aircraft, Cars, and 5Datasets. On CUB-200 and Flowers, although CL-LoRA remains stronger, FS-EWCLoRA consistently improves over its matched EWCLoRA baseline. Overall, the trajectories indicate that the improvements are generally accumulated during sequential learning rather than arising only at the final stage.

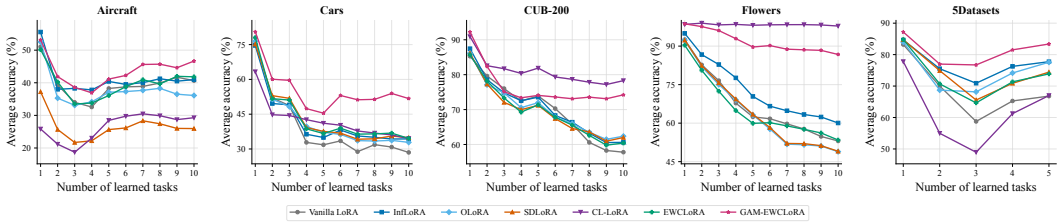


Figure 6: Stage-wise average-accuracy trajectories on the fine-grained and heterogeneous vision benchmarks reported in Table 2.

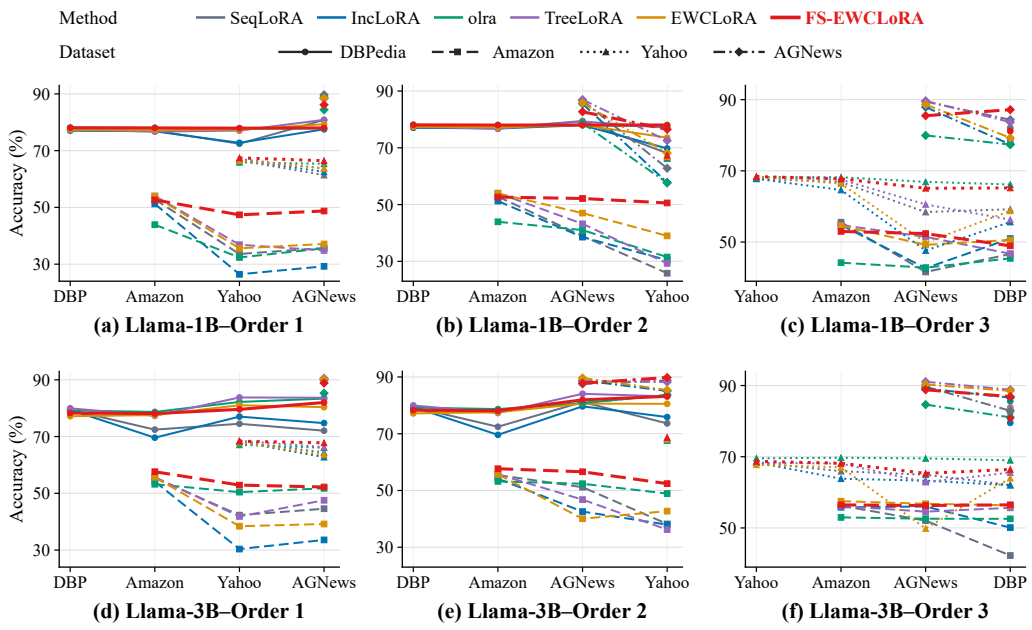


Figure 7: Task-wise accuracy trajectories for Llama-3.2-1B and Llama-3.2-3B across three continual text-classification task orders. Each curve starts when the corresponding dataset is learned and tracks its accuracy after subsequent tasks.

Stage-wise performance on continual NLP benchmark. Fig. 7 complements Table 3 by showing how individual task accuracies evolve under subsequent learning. Across most orders, FS-EWCLoRA better preserves previously learned tasks, particularly those exposed to multiple later updates, while the gains are not necessarily uniform across every dataset. The main exception is Llama-3.2-3B under Order 3, where weaker final performance on some tasks offsets improved retention on others and produces a slightly lower FAA than EWCLoRA. Overall, the trajectories support the more consistent performance of FS-EWCLoRA across the evaluated task orders.

B.3 Additional Two-Task Solution-Shaping Diagnostics

Table 10 shows a consistent pattern across both parameterizations. Under the same $\mathsf{C}+\mathsf{P}$ objective on Task 2, shaping the inherited Task 1 solution increases its retained accuracy from 87.95 to 91.82 under LoRA and from 85.53 to 87.89 under full-parameter fine-tuning. The complete FS-EWC configuration also yields higher final Task 1 accuracy than EWC in both settings, while improving Task 2 accuracy under LoRA and leaving it nearly unchanged under full fine-tuning. For LoRA, the more negative BWT of FS-EWC reflects its higher Task 1 starting accuracy rather than lower final retention. These results reproduce the main mechanism across parameterizations: subsequent retention depends on the solution passed forward as well as on the protection subsequently applied to it.

Table 10: Two-task solution-shaping diagnostics under LoRA adaptation and full-parameter fine-tuning on ImageNet-R with ViT-B/16. C, F, and P denote current-task fitting, Fisher shaping, and historical EWC protection, respectively. $\mathcal{F}_1$ denotes the Task-1 Fisher statistic reported under each parameterization, and d_2 denotes the Task-2 parameter displacement in the corresponding trainable parameter space. All variants use endpoint references without trajectory averaging.

Model	Variant	Training objective		Task 1 endpoint		Task 2 endpoint			Task 2 update
		Task 1	Task 2	$a_{11}\uparrow$	$\mathcal{F}_1$	$a_{21}\uparrow$	$a_{22}\uparrow$	BWT↑	d_2
LoRA	Vanilla	C	C	90.73	19.885	84.85	89.59	-5.87	1.560
	Shaping only	$\mathsf{C}+\mathsf{F}$	$\mathsf{C}+\mathsf{F}$	93.34	10.946	84.80	93.81	-8.55	3.131
	EWC	C	$\mathsf{C}+\mathsf{P}$	90.73	19.885	87.95	83.45	-2.78	0.0106
	FS-EWC	$\mathsf{C}+\mathsf{F}$	$\mathsf{C}+\mathsf{F}+\mathsf{P}$	93.34	10.946	89.10	88.33	-4.24	0.0204
	Task-1 shaping	$\mathsf{C}+\mathsf{F}$	$\mathsf{C}+\mathsf{P}$	93.34	10.946	91.82	82.57	-1.52	0.392
	Task-2 shaping	C	$\mathsf{C}+\mathsf{F}+\mathsf{P}$	90.73	19.885	83.07	89.21	-7.66	0.129
Full FT	Vanilla	C	C	92.61	869.80	78.46	92.47	-14.15	4.941
	Shaping only	$\mathsf{C}+\mathsf{F}$	$\mathsf{C}+\mathsf{F}$	93.24	858.59	81.92	93.44	-11.32	5.131
	EWC	C	$\mathsf{C}+\mathsf{P}$	92.61	869.80	85.53	90.66	-7.08	4.747
	FS-EWC	$\mathsf{C}+\mathsf{F}$	$\mathsf{C}+\mathsf{F}+\mathsf{P}$	93.24	858.59	90.25	90.79	-2.99	4.673
	Task-1 shaping	$\mathsf{C}+\mathsf{F}$	$\mathsf{C}+\mathsf{P}$	93.24	858.59	87.89	87.31	-5.35	4.774
	Task-2 shaping	C	$\mathsf{C}+\mathsf{F}+\mathsf{P}$	92.61	869.80	82.08	93.86	-10.53	4.595

B.4 Complementarity Study: Experimental Details and Sensitivity to λ

We further evaluate the Fisher-complementarity study over three random seeds. For all Fisher computations, we keep the historical regularization coefficient fixed across methods and seeds, using $\lambda = 2 \times 10^6$ for the diagonal representation and $\lambda = 2 \times 10^5$ for the Kronecker-factored representation. These coefficients are also shared across the batched empirical, per-example empirical, and model-sampled Fisher variants. Unless otherwise specified, all remaining training and retained-state construction settings follow Table 5.

Table 11: Multi-seed results for the Fisher-complementarity study in Table 5. FAA and AAA are reported as mean $\pm$ sample standard deviation (%) over seeds 0, 1, and 2.

Method	Metric	Diagonal			Kronecker-factored		
		Batched	Per-example	Model-sampled	Batched	Per-example	Model-sampled
EWC	FAA	$51.96_{(1.15)}$	$50.60_{(1.07)}$	$50.76_{(1.14)}$	$53.74_{(0.47)}$	$52.75_{(0.78)}$	$52.71_{(1.17)}$
	AAA	$66.30_{(0.53)}$	$66.60_{(0.97)}$	$66.60_{(1.09)}$	$67.37_{(0.18)}$	$67.55_{(0.94)}$	$67.33_{(0.89)}$
FS-EWC	FAA	$59.29_{(0.55)}$	$59.67_{(1.07)}$	$59.62_{(0.88)}$	$59.82_{(0.91)}$	$59.95_{(0.93)}$	$59.62_{(0.82)}$
	AAA	$70.69_{(0.67)}$	$70.88_{(0.50)}$	$70.86_{(0.51)}$	$71.00_{(0.78)}$	$71.16_{(0.47)}$	$71.12_{(0.46)}$

Multi-seed consistency. Table 11 shows that FS-EWC improves over the matched EWC baseline under all six combinations of Fisher computation and matrix representation. The improvements range from 6.08 to 9.07 percentage points in FAA and from 3.61 to 4.39 points in AAA. Kronecker-factored Fisher generally provides modest gains over the diagonal representation, particularly for EWC, while the advantage of FS-EWC persists under both representations and all three Fisher computations. These results confirm that the complementarity observed in Table 5 is consistent across random seeds.

Sensitivity to historical regularization strength. Table 12 further examines the historical regularization coefficient λ under batched empirical Fisher. Across both diagonal and Kronecker-factored representations, FS-EWC remains above the corresponding EWC baseline at every numerically valid matched value of λ. Its performance also varies relatively little over the evaluated range, whereas EWC is more sensitive to the strength of historical regularization. This indicates that the gain from solution shaping is not confined to a narrowly selected protection coefficient.

Table 12: Sensitivity to the historical regularization coefficient λ under batched empirical Fisher. Each entry reports FAA / AAA (%) for seed 2.

Representation	Method	Historical regularization coefficient λ					
		0	2×10^2	2×10^3	2×10^4	2×10^5	2×10^6
Diagonal	EWC	49.51 / 64.95	49.24 / 64.66	49.37 / 64.66	49.70 / 65.01	50.56 / 65.55	52.07 / 66.38
	FS-EWC	58.53 / 69.82	58.79 / 69.65	58.39 / 69.75	58.59 / 69.66	58.39 / 69.73	58.66 / 69.94
Kronecker-factored	EWC	49.31 / 64.92	49.31 / 64.89	49.70 / 65.15	50.23 / 66.15	53.32 / 67.33	NaN
	FS-EWC	58.85 / 69.80	58.39 / 69.78	58.66 / 69.80	58.46 / 69.81	58.79 / 70.14	58.92 / 70.57

The Kronecker-factored EWC run at $\lambda = 2 \times 10^6$ encountered numerical instability and produced NaN values.

B.5 Retained-State Construction Details

Table 13 shows that trajectory averaging achieves the highest AAA for all three seeds, whereas denser trajectory sampling provides no further improvement. All multi-checkpoint constructions yield less negative BWT than the endpoint reference, although endpoint-plus-initial achieves the least negative BWT at the cost of substantially lower FAA and AAA. Including the initialization provides no additional accuracy benefit for either endpoint or trajectory averaging. Overall, these results indicate that learned states along the training trajectory provide useful information for retained-state construction, but neither denser sampling nor simply averaging more states is inherently beneficial.

Table 13: Multi-seed results for retained-state construction on ImageNet-A. Each retained state is constructed by uniformly averaging the LoRA factors at checkpoints $\mathcal{K}_t$, with $\omega = 1$; the current-task Fisher is estimated after merging. Checkpoint 0 denotes the state before the first optimizer update of each task, while checkpoint 50 denotes the terminal iterate.

Retained state	Checkpoints $\mathcal{K}_t$	Metric	Seed 0	Seed 1	Seed 2	Mean $\pm$ SD
Endpoint	$\{50\}$	FAA	58.20	59.97	59.51	$59.23_{(0.92)}$
		AAA	70.48	71.00	70.41	$70.63_{(0.33)}$
		BWT	-7.29	-8.66	-8.29	$-8.08_{(0.71)}$
Late-stage	$\{30, 40, 50\}$	FAA	58.53	59.84	59.91	$59.43_{(0.78)}$
		AAA	70.51	70.98	70.35	$70.61_{(0.33)}$
		BWT	-7.01	-8.35	-7.68	$-7.68_{(0.67)}$
Trajectory	$\{10, 20, 30, 40, 50\}$	FAA	58.59	59.78	59.64	$59.34_{(0.65)}$
		AAA	70.63	71.29	70.61	$70.84_{(0.39)}$
		BWT	-6.89	-7.70	-7.81	$-7.47_{(0.50)}$
Dense trajectory	$\{5, 10, \ldots, 50\}$	FAA	58.39	59.64	59.91	$59.31_{(0.81)}$
		AAA	70.55	71.15	70.52	$70.74_{(0.35)}$
		BWT	-7.00	-8.01	-7.32	$-7.44_{(0.52)}$
Endpoint + initial	$\{0, 50\}$	FAA	56.88	58.92	59.51	$58.44_{(1.38)}$
		AAA	68.87	69.72	69.56	$69.38_{(0.45)}$
		BWT	-6.52	-6.34	-6.44	$-6.44_{(0.09)}$
Trajectory + initial	$\{0, 10, 20, 30, 40, 50\}$	FAA	57.93	59.38	59.71	$59.01_{(0.95)}$
		AAA	70.32	70.94	70.49	$70.59_{(0.32)}$
		BWT	-6.84	-7.58	-7.57	$-7.33_{(0.43)}$