BIFILT: BILATERAL SUBSPACE FILTERING FOR DATA-FREE CONTINUAL MODEL MERGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Data-free continual model merging (DFCMM) aims to sequentially integrate independently fine-tuned experts into a single model without access to task data or previously seen experts. Existing subspace-based methods typically protect historical knowledge through input-side filtering, imposing $ZV = 0$ and thereby suppressing all update responses along selected historical input directions. We show that preserving the selected historical input-output core requires only the weaker condition $U^\top ZV = 0$, revealing cross-subspace updates that are removed by one-sided filtering despite leaving the protected core unchanged. Based on this observation, we introduce BiFilt, a bilateral subspace filtering framework that decomposes each incoming update into four input-output interaction blocks, removes the jointly historical component, preserves the complementary component, and independently controls the two directional cross components. We derive BiFilt as the unique solution of a convex quadratic objective, yielding a simple closed-form filter and a fixed symmetric configuration without task-dependent tuning. Controlled task-pair analyses show that the released cross direction contains substantial new-task information while causing only limited historical degradation, and complete continual-merging experiments across multiple CLIP backbones and task horizons consistently demonstrate the benefit of partial cross-block retention over one-sided filtering. Further analyses reveal a stability-plasticity trade-off: partial retention improves task acquisition and final accuracy on standard sequences, whereas stricter filtering can provide stronger retention over longer horizons. These results show that distinguishing input-output interaction directions provides a more flexible parameter-only principle for continual model merging.

1 INTRODUCTION

Model merging combines independently fine-tuned models into a single model without joint retraining or simultaneous access to their original training data (Wortsman et al., 2022; Ilharco et al., 2023; Yadav et al., 2023). Data-free continual model merging (DFCMM) further makes this setting sequential: task experts arrive one at a time, and each merge has access only to the common pre-trained model, the current merged model, and the incoming expert (Tang et al., 2025; Yang et al., 2025; Qiu et al., 2026b). The central challenge is therefore a stability–plasticity trade-off: an incoming update should introduce useful task-specific behavior while limiting interference with capabilities accumulated from previous experts.

Recent DFCMM methods use parameter-space subspaces as data-free proxies for historical knowledge (Tang et al., 2025; Yang et al., 2025; Qiu et al., 2026b). In particular, NUFILT protects selected historical input directions V by requiring the filtered increment Z to satisfy

$$ZV = 0, \tag{1}$$

thereby suppressing every new response along those directions (Qiu et al., 2026b). This provides a clear protection mechanism, but makes the filtering decision entirely from the *input side*. A matrix update, however, has both an input and an output direction. Two updates may act on the same historical input direction while producing responses in fundamentally different output subspaces. Treating them identically can therefore remove updates that interact with historical inputs without modifying the historical output coordinates that we actually seek to preserve.

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.

This distinction becomes explicit when both singular subspaces of the accumulated update are considered. Let U and V denote its selected historical output and input subspaces, respectively. Preserving their compressed input–output map requires only

$$U^\top ZV = 0, \tag{2}$$

rather than the stronger condition $ZV = 0$. With the left and right projectors induced by U and V, an incoming update τ admits the orthogonal decomposition

$$\tau = \tau_{\parallel\parallel} + \tau_{\parallel\perp} + \tau_{\perp\parallel} + \tau_{\perp\perp}. \tag{3}$$

Right-sided filtering removes both $\tau_{\parallel\parallel}$ and $\tau_{\perp\parallel}$ because they share historical input directions, even though only the former lies simultaneously in the selected historical input and output subspaces. The latter is a *cross interaction*: it acts on historical inputs while producing responses in complementary output directions. Symmetrically, $\tau_{\parallel\perp}$ captures the opposite cross interaction. This four-block view exposes a degree of freedom that one-sided filtering cannot express: the jointly historical component can be removed while the two directional cross components are controlled separately.

Based on this observation, we introduce BiFilt, a bilateral subspace filtering framework for data-free continual model merging. BiFilt removes the jointly historical component $\tau_{\parallel\parallel}$, preserves the doubly complementary component $\tau_{\perp\perp}$, and independently shrinks the two directional cross components:

$$\mathcal{F}_{\alpha,\beta}(\tau) = \tau_{\perp\perp} + \beta\tau_{\parallel\perp} + \alpha\tau_{\perp\parallel}. \tag{4}$$

Rather than introducing α and β heuristically, we derive this operator as the unique solution of a convex quadratic objective that balances fidelity to the incoming update with penalties on responses involving the historical output and input subspaces. Right-sided, left-sided, and hard bilateral filtering arise naturally as boundary cases. Assigning equal unit penalties to the two historical sides gives the fixed symmetric configuration $\alpha = \beta = 0.5$, which we use throughout the main evaluation without task-dependent coefficient selection. A subsequent finite-rank recovery stage fits historical and incoming response targets. Accordingly, the core-preservation guarantee applies exactly to the filtered proposal, while recovery may redistribute its final block geometry.

The block decomposition is not merely algebraic. Across 56 ordered task pairs, the complementary-output/historical-input block $\tau_{\perp\parallel}$ accounts for 26.9% of the incoming-update energy on average. Restoring this component improves current-task accuracy in all 56 pairs, by 5.2 percentage points on average, while changing historical-task accuracy by only -0.2 points. These observations indicate that components discarded solely because of historical input alignment can contain substantial new-task information. Complete continual sequences further reveal the corresponding stability–plasticity trade-off: partial cross-block retention improves task acquisition and final accuracy on standard eight-task sequences, whereas stricter filtering can provide stronger final retention as the task horizon grows. Moreover, the two cross directions exhibit different empirical behavior, supporting their explicit rather than interchangeable treatment.

Our main contributions are:

- **A more precise preservation principle.** We distinguish suppressing all responses on selected historical inputs, $ZV = 0$, from preserving the selected historical input–output core, $U^\top ZV = 0$, and show that the former imposes a strictly stronger constraint.
- **A bilateral filtering formulation.** We decompose incoming updates into four directional interaction blocks and derive a unique closed-form filter that removes joint historical overlap while independently controlling the two cross interactions. Existing one-sided and hard bilateral filtering rules emerge as boundary cases.
- **Mechanistic and continual-merging evidence.** Ordered-task-pair experiments show that a cross component removed by right-sided filtering carries substantial new-task information, while controlled continual experiments isolate the effects of cross-block retention, recovery, historical rank, and task horizon. The results reveal a consistent stability–plasticity trade-off rather than a universally optimal filtering strength.

2 Related Work

Model merging and interference mitigation. Model merging combines independently fine-tuned models without joint retraining. Representative model merging approaches include parameter averag-

ing (Wortsman et al., 2022), Fisher-weighted merging (Matena & Raffel, 2022), task arithmetic (Ilharco et al., 2023), and TIES (Yadav et al., 2023). AdaMerging adapts task-vector coefficients using unlabeled test data (Yang et al., 2024), while WUDI-Merging uses task-vector geometry to suppress interference in a data-free setting (Cheng et al., 2025). MaTS provides a broader optimization view by matching models within task parameter subspaces (Tam et al., 2024). These methods mainly study simultaneous or offline merging, whereas BiFilt focuses on the sequential setting in which only the current merged model and one incoming expert are available.

Low-rank and subspace structure. Recent work increasingly exploits matrix structure in task updates. KnOTS aligns LoRA updates through SVD before merging (Stoica et al., 2025), No Task Left Behind separates common and task-specific singular subspaces (Marczak et al., 2025), and AdaRank selects useful singular components using unlabeled data (Lee et al., 2026). SMILE and MASS instead retain task-specific low-rank components and combine them through sparse expert selection or inference-time routing (Tang et al., 2026; Crisostomi et al., 2026). BiFilt uses subspaces differently: the left and right singular spaces of the accumulated update define historical output and input coordinates, whose joint projector structure decomposes an incoming update into four directional interaction blocks before fusion. The resulting model remains a single backbone without an additional inference-time router.

Data-free continual model merging. Continual model merging processes independently trained experts sequentially. MagMax first connected merging to continual learning by combining sequential fine-tuning with maximum-magnitude weight selection (Marczak et al., 2024). OPCM regulates incoming task vectors using accumulated parameter directions (Tang et al., 2025), while DOP balances stability and plasticity through dual projection objectives (Yang et al., 2025). NUFILT is the closest setting to ours: it projects incoming updates away from selected historical input directions and then applies low-rank recovery (Qiu et al., 2026b). SAFE-Merge further considers preservation of pre-trained general knowledge through sparse masking and constrained low-rank recovery (Qiu et al., 2026a).

Related continual-learning methods also identify low-interference directions through representation or functional subspaces. GPM and Adam-NSCL construct protected spaces from historical activations or feature statistics (Saha et al., 2021; Wang et al., 2021), while OGD explicitly projects new updates into directions that preserve previous network outputs to first order (Farajtabar et al., 2020). Jacobian-based regularization has likewise been used to preserve previous tasks in continual learning (Heckel, 2022). BiFilt differs in that it has no access to historical task data or Jacobians: both sides of its historical geometry are derived solely from accumulated parameter updates.

3 Motivation: Beyond Input-Side Suppression

For selected historical singular spaces U, V of the accumulated update S, preserving the compressed map $U^\top SV$ requires

$$U^\top ZV = 0.$$

The right-sided rule $ZV = 0$ is stronger. For a selected historical input direction $v \in \mathrm{span}(V)$, the vector $\tau_t v$ denotes the output response induced by the incoming expert update on that direction. It decomposes as

$$\tau_t v = P_U \tau_t v + P_U^\perp \tau_t v,$$

where the first term lies in the selected historical output subspace and the second lies in its complement. Right-sided filtering suppresses the entire response, whereas preserving the selected core only requires the historical-output component to vanish.

Indeed, for a unit vector $v \in \mathrm{span}(V)$ and nonzero $q \in \mathrm{span}(U)^\perp$, the update $Z = qv^\top$ satisfies

$$Zv = q \neq 0, \qquad U^\top ZV = 0.$$

It can therefore produce a nonzero response on a historical input while remaining invisible to the selected historical output coordinates. Figure 1 illustrates this distinction.

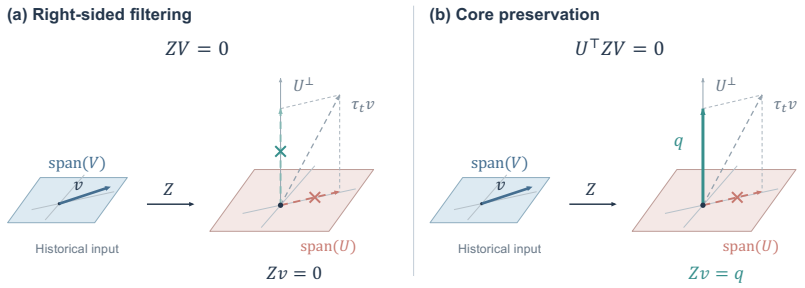

Figure 1: **Core preservation versus right-sided filtering.** For $v \in \mathrm{span}(V)$, the right-sided rule sets $Zv = 0$; core preservation permits $Zv = q \in \mathrm{span}(U)^\perp$.

A functional safety view. Let $\mathcal{H}_\ell$ and $\mathcal{G}_\ell$ denote the spans of historical layer inputs and output-side loss gradients, respectively, and define $\mathfrak{S}_\ell = \{Z_\ell : g^\top Z_\ell h = 0, \forall h \in \mathcal{H}_\ell, \forall g \in \mathcal{G}_\ell\}$.

Theorem 1 (First-order safe-space characterization)**.** *With orthogonal projectors $P_{\mathcal{H}_\ell}$ and $P_{\mathcal{G}_\ell}$,*

$$\mathfrak{S}_\ell = \{Z_\ell : P_{\mathcal{G}_\ell} Z_\ell P_{\mathcal{H}_\ell} = 0\} = \{Z_\ell : Z_\ell \mathcal{H}_\ell \subseteq \mathcal{G}_\ell^\perp\}.$$

If $\mathcal{H}_\ell \subseteq \mathrm{span}(V_\ell)$ and $\mathcal{G}_\ell \subseteq \mathrm{span}(U_\ell)$, then $U_\ell^\top Z_\ell V_\ell = 0$ is a sufficient parameter-only condition for $Z_\ell \in \mathfrak{S}_\ell$. Proofs and the downstream extension are given in Appendix A.1.

The geometric argument only identifies directions that are admissible under the selected-core constraint; it does not guarantee that they are useful. We therefore test whether the released complementary-output/historical-input (CH) direction carries incoming-task signal without inducing a comparably large historical cost.

Across 56 ordered task pairs, the CH ($\perp\|$) block contains 26.9% of incoming-update energy on average. Restoring it raises current-task accuracy in all 56 pairs, with a mean gain of 5.2 pp, while historical-task accuracy changes by only -0.2 pp on average. CH energy share is descriptively associated with the current-task gain ($r = 0.6$; Figure 2). The ordered pairs are used as descriptive observations rather than independent replicates. The consistent current-task gain together with the small but nonzero historical effect motivates partial rather than unrestricted cross-block retention.

The same input–output factorization also exposes the opposite historical-output/complementary-input (HC) interaction. This motivates a bilateral formulation that keeps the two cross directions separately controllable rather than treating input-side alignment alone as the filtering criterion.

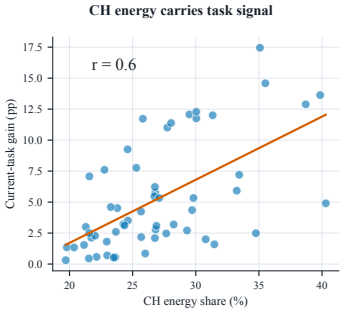

Figure 2: **The released CH direction carries new-task signal.** CH ($\perp\|$) share versus the current-task accuracy gain from restoring this component across 56 ordered task pairs. The fitted line shows a descriptive association ($r = 0.6$); task pairs are not treated as independent replicates.

4 Method

4.1 Problem Setup

Let θ_0 be the shared pre-trained model and θ_t the expert arriving at stage t. Only θ_0, the previous merged model, and θ_t are available. The conceptual risk averages task losses over observed tasks, but task data and earlier experts are unavailable at merge time. For each selected matrix layer, define the historical and incoming updates by

$$S_{t-1} = W_{t-1}^{\mathrm{merged}} - W_0, \qquad \tau_t = W_t - W_0. \tag{5}$$

For a fixed stage, we write $S = S_{t-1}$ and $\tau = \tau_t$ when no stage index is needed, and suppress layer indices below. At stage 1, $S_0 = 0$ has no unique nontrivial historical subspace; an explicit first-stage

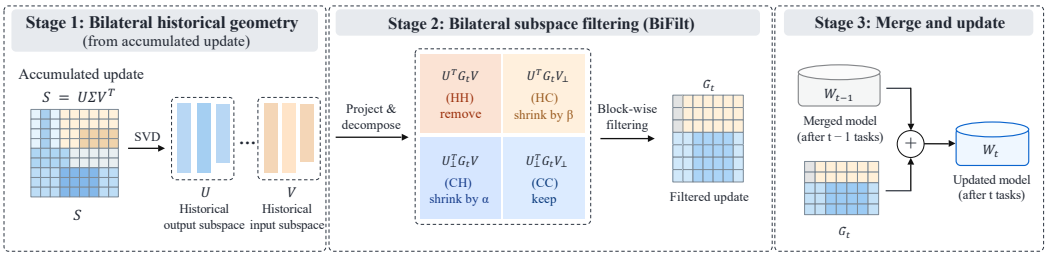

Figure 3: **Overview of BiFilt.** The accumulated update provides historical output and input subspaces U and V. Their joint projector structure decomposes an incoming update into four interaction blocks: jointly historical (HH, $\|\|$), historical-output/complementary-input (HC, $\|\perp$), complementary-output/historical-input (CH, $\perp\|$), and doubly complementary (CC, $\perp\perp$). BiFilt removes HH, scales HC and CH by β and α, respectively, and preserves CC. Section 4.4 describes the subsequent recovery and fusion stage.

rule $\mathcal{B}$ is therefore part of each experimental protocol, and recovery is skipped. For $t \geq 2$, BiFilt filters and recovers selected layers; other parameters follow the common baseline rule. All temporary bases and recovery factors are discarded after fusion.

4.2 Preserving the Historical Input–Output Core

Let U, V be orthonormal leading left and right singular bases of S, with projectors $P_U = UU^\top$ and $P_V = VV^\top$. Preserving the selected core $U^\top SV$ under increment Z is equivalent to

$$U^\top ZV = 0 \quad \Longleftrightarrow \quad P_U Z P_V = 0. \tag{6}$$

When the historical input space and complementary output space are nontrivial,

$$\{Z : ZV = 0\} \subsetneq \{Z : U^\top ZV = 0\}.$$

Under the coverage condition of Theorem 1, $\mathcal{A} = \{Z : U^\top ZV = 0\}$ is a parameter-only sufficient subset of the functional safe space. Its Frobenius projection of τ is $\tau - P_U \tau P_V$ (Corollary 2), which removes only HH. BiFilt further regularizes the two cross interactions:

$$\min_{Z:\, P_U Z P_V = 0} \ \tfrac{1}{2}\|Z - \tau\|_F^2 + \tfrac{\lambda_U}{2}\|P_U Z\|_F^2 + \tfrac{\lambda_V}{2}\|Z P_V\|_F^2, \quad \lambda_U, \lambda_V \geq 0. \tag{7}$$

The penalties separately regulate historical output and input responses; their connection to task behavior remains conditional on the alignment discussed in Section 3.

4.3 Closed-Form Bilateral Filtering

The projectors split τ into four orthogonal blocks $\tau_{ab} = P_U^a \tau P_V^b$, where $a, b \in \{\|, \perp\}$, $P_U^{\|} = P_U$, and $P_U^{\perp} = I - P_U$ (similarly for V). The first index refers to output alignment. The core constraint removes HH ($\|\|$); the objective retains CC ($\perp\perp$) and gives the unique solution

$$\widehat{\tau} = \mathcal{F}_{\alpha,\beta}(\tau) = \tau_{\perp\perp} + \beta\tau_{\|\perp} + \alpha\tau_{\perp\|}, \qquad \beta = (1 + \lambda_U)^{-1}, \quad \alpha = (1 + \lambda_V)^{-1}. \tag{8}$$

Thus HC (historical output only) is scaled by β, and CH (historical input only) by α. Right-sided, left-sided, and hard bilateral filtering correspond to $(\alpha, \beta) = (0, 1), (1, 0), (0, 0)$. Equal unit penalties give the fixed BiFilt-Sym rule $(\alpha, \beta) = (0.5, 0.5)$, the mean of the two one-sided hard filters. Unlike a product of independent soft projectors, BiFilt removes HH exactly. Appendix A.2–A.5 gives the decomposition, boundary identities, and derivation.

4.4 Projection-Aware Recovery and Fusion

Following NUFILT (Qiu et al., 2026b), recovery fits historical and incoming responses using the rank-r correction

$$R_t = \tau_t B_t A_t, \quad B_t \in \mathbb{R}^{d_{\text{in}} \times r}, \quad A_t \in \mathbb{R}^{r \times d_{\text{in}}}. \tag{9}$$

Algorithm 1 BiFilt procedure

Require: Pre-trained model θ_0; arriving experts $\theta_1, \ldots, \theta_T$; selected layers $\mathcal{J}$; historical rank k; probe ranks $k_{\mathrm{h}}, k_{\mathrm{n}}$; recovery rank r; retention coefficients α, β; first-stage rule $\mathcal{B}$; recovery initialization $\mathcal{I}$; optimizer and budget $\mathrm{MaxIter}$; rule $\mathcal{G}$ for remaining parameters

$\theta_1^{\mathrm{merged}} \leftarrow \mathcal{B}(\theta_0, \theta_1)$ ▷ First-stage protocol; no recovery
for $t = 2, \ldots, T$ **do**
  Treat $\theta_{t-1}^{\mathrm{merged}}$ as a fixed snapshot
  for each matrix layer $\ell \in \mathcal{J}$ **do**
    $S_{t-1}^{(\ell)} \leftarrow (W_{t-1}^{\mathrm{merged}})^{(\ell)} - W_0^{(\ell)}$
    $\tau_t^{(\ell)} \leftarrow W_t^{(\ell)} - W_0^{(\ell)}$
    Compute rank-k historical bases $U^{(\ell)}, V^{(\ell)}$ from $S_{t-1}^{(\ell)}$
    $\widehat{\tau}_t^{(\ell)} \leftarrow \mathcal{F}_{\alpha,\beta}^{(\ell)}(\tau_t^{(\ell)})$ using Eq. 8
    Construct $\widehat{V}^{(\ell)}$ and $Y^{(\ell)}$ using Eq. 10
    Initialize $(A_t^{(\ell)}, B_t^{(\ell)}) \leftarrow \mathcal{I}$
  end for
  Optimize the sum of Eq. 11 over $\ell \in \mathcal{J}$ for $\mathrm{MaxIter}$ steps
  for each matrix layer $\ell \in \mathcal{J}$ **do**
    $(W_t^{\mathrm{merged}})^{(\ell)} \leftarrow (W_{t-1}^{\mathrm{merged}})^{(\ell)} + \widehat{\tau}_t^{(\ell)} + \tau_t^{(\ell)} B_t^{(\ell)} A_t^{(\ell)}$
  end for
  Process parameters outside $\mathcal{J}$ using $\mathcal{G}(\theta_0, \theta_{t-1}^{\mathrm{merged}}, \theta_t)$
  Discard the incoming expert, singular bases, and recovery factors
end for
return $\theta_T^{\mathrm{merged}}$

For historical and incoming right-singular probes $\widehat{V}_{\le t-1}$ and $\widehat{V}_t$, set $\widehat{V} = [\widehat{V}_{\le t-1}, \widehat{V}_t]$ and

$$Y = [S_{t-1}\widehat{V}_{\le t-1},\ \tau_t \widehat{V}_t]. \tag{10}$$

With $M = S_{t-1} + \widehat{\tau}_t$, the factors minimize the sum over selected layers of

$$\mathcal{L}(A_t, B_t) = \|Y - (M + \tau_t B_t A_t)\widehat{V}\|_F^2. \tag{11}$$

The fused increment is $\widetilde{\tau}_t = \widehat{\tau}_t + R_t$. Although $P_U \widehat{\tau}_t P_V = 0$, generally $P_U \widetilde{\tau}_t P_V = P_U R_t P_V \neq 0$; exact core preservation therefore applies to the pre-recovery proposal. Recovery factors are folded into the merged weights and introduce no additional inference module.

Algorithm 1 gives the complete continual-merging procedure, including first-stage initialization, layerwise bilateral filtering, recovery optimization, fusion, and the treatment of remaining parameters. The recovery-capacity and complexity analyses are deferred to Appendices A.6 and A.8.

5 Experiments

5.1 Experimental Setup

We evaluate shared-initialization CLIP visual experts (Radford et al., 2021) on three backbones (ViT-B/32, ViT-B/16, and ViT-L/14) and 8-, 14-, and 20-task suites. Methods within each setting use matched task orders and full-test-set evaluation after every merge stage.

First-stage protocols. Because $S_0 = 0$ defines no unique historical subspace, we treat the first-stage rule as an explicit protocol choice. The broad benchmark retains the established $\mathcal{B}_{\mathrm{SVD}}$ convention for comparability, whereas structural analyses use the audited rank-zero rule $\mathcal{B}_0$. A matched diagnostic with three independently generated orthogonal bases on ViT-B/32 TA8 yields a consistent $+0.3$ pp ACC effect with negligible mean BWT change. We therefore keep the two archived protocol cohorts separate; full protocol definitions and the matched diagnostic are provided in Appendices B and C.1.

Table 1: **Broad continual-merging comparison under the established benchmark protocol.** ACC (%) and BWT (pp), higher is better. Results are mean ± sample SD over ten matched task orders (seeds 42–51). The $\mathcal{B}_{\mathrm{SVD}}$ convention retains the established first-stage behavior of the continual-merging benchmark pipeline; expert checkpoints, task orders, test sets, and evaluation procedures are matched within each setting. The benchmark is reported separately from the audited $\mathcal{B}_0$ structural cohort. **Bold** marks the best merging-method mean per column, including ties.

	Method	ViT-B/32			ViT-B/16			ViT-L/14		
		8 tasks	14 tasks	20 tasks	8 tasks	14 tasks	20 tasks	8 tasks	14 tasks	20 tasks
	PRE-TRAINED	48.1	56.9	55.6	55.4	62.0	59.8	64.9	69.1	65.6
	INDIVIDUAL	90.4	89.3	89.8	92.4	91.3	91.6	94.3	93.4	93.5
	C. FINE-TUNED	79.8	67.4	62.6	82.9	72.2	68.2	90.0	70.9	77.7
ACC (%) ↑	WEIGHT AVERAGING	66.3±0.0	65.4±0.0	61.1±0.0	72.3±0.0	69.7±0.0	64.8±0.0	80.0±0.0	77.5±0.0	71.1±0.0
	TASK ARITHMETIC	67.5±0.0	66.5±0.0	60.0±0.0	77.1±0.0	70.9±0.6	64.2±0.0	82.1±0.0	77.9±0.0	70.3±0.0
	TIES-MERGING	49.0±10.2	66.2±0.6	59.9±0.7	66.8±3.7	70.5±0.8	63.0±1.6	64.3±7.0	78.0±0.6	68.3±0.9
	MAGMAX-IND	70.7±0.0	67.0±0.0	61.2±0.0	76.7±1.8	67.0±0.0	62.5±0.0	83.4±0.0	71.2±0.0	71.2±0.0
	DOP	78.3±1.6	73.0±1.2	68.5±1.6	84.5±0.8	76.2±1.2	72.2±1.8	88.3±0.7	85.5±0.8	80.7±0.9
	WUDI-MERGING	74.7±6.6	67.0±6.9	63.7±3.8	81.0±4.7	75.0±4.1	69.6±4.7	87.5±3.3	84.2±3.7	78.1±2.8
	OPCM	75.5±0.5	71.9±0.3	65.7±0.2	81.8±0.3	77.1±0.5	70.3±0.2	87.0±0.4	83.5±0.2	76.0±0.2
	NUFILT	83.6±0.2	78.0±0.2	71.0±0.9	87.3±0.1	83.1±0.3	78.1±0.9	91.6±0.1	89.2±0.1	84.7±0.8
	BiFilt-Sym (Ours)	**84.0**±0.2	**79.2**±0.2	**72.6**±0.8	**87.5**±0.1	**84.0**±0.3	**79.6**±0.8	**91.8**±0.1	**89.6**±0.1	**85.5**±0.5
	WEIGHT AVERAGING	-11.5±2.2	-8.0±1.3	-7.1±2.1	-9.7±1.5	-7.1±1.4	-7.3±1.7	-7.3±1.4	-5.8±1.0	-6.4±1.5
	TASK ARITHMETIC	-9.6±1.5	-1.3±1.6	-3.4±1.0	-4.2±1.0	-1.3±0.4	-3.6±0.4	-7.1±0.8	-1.8±0.3	-3.3±0.3
	TIES-MERGING	-15.3±8.0	**1.9**±0.6	**-1.5**±0.7	-5.5±0.4	**1.4**±0.7	**-1.5**±1.2	-13.0±5.7	**-1.1**±0.4	**-2.9**±1.0
	MAGMAX-IND	-8.3±1.3	-7.4±1.4	-7.2±1.6	-6.1±1.3	-7.4±2.0	-8.0±2.2	-5.0±0.8	-6.0±2.1	-6.5±2.1
	DOP	-6.1±2.3	-8.8±2.1	-12.5±2.8	-3.9±1.0	-7.5±2.2	-11.4±2.8	-3.2±1.2	-4.6±1.0	-8.4±1.4
	WUDI-MERGING	-17.0±7.5	-22.8±7.3	-26.0±4.1	-12.6±5.4	-16.9±4.4	-18.5±14.2	-7.3±3.7	-9.4±4.0	-15.8±2.9
BWT (pp) ↑	OPCM	-6.3±1.1	-6.0±1.0	-7.8±1.5	-4.8±0.7	-5.1±1.4	-6.3±2.2	-2.6±1.0	-4.3±0.7	-6.5±1.8
	NUFILT	-2.7±0.7	-5.7±0.9	-8.9±2.3	-1.6±0.5	-3.5±0.6	-7.1±1.9	-1.1±0.3	-2.0±0.3	-4.6±0.7
	BiFilt-Sym (Ours)	**-2.2**±0.6	-4.9±0.7	-8.3±2.0	**-1.3**±0.4	-3.0±0.6	-6.2±1.6	**-0.9**±0.3	-1.8±0.3	-4.4±0.6

Filtering and recovery configuration. Unless stated otherwise, structural controls use historical rank 128, incoming-probe rank 8, rank-64 recovery, 50 Adam steps, unit fusion scale, and fixed $(\alpha, \beta) = (0.5, 0.5)$. ViT-B/32 TA8 structural comparisons use seeds 42, 43, 44, 46, and 47; ViT-B/16 and the B/32 design controls use seeds 43, 44, and 46. A complete seeded sequence is the statistical unit.

For task i after stage t, let $a_{t,i}$ be accuracy. Following standard continual-learning evaluation (Lopez-Paz & Ranzato, 2017), we report final $\mathrm{ACC} = T^{-1}\sum_i a_{T,i}$ and backward transfer $\mathrm{BWT} = (T-1)^{-1}\sum_{i<T}(a_{T,i} - a_{i,i})$. Task-arrival $\mathrm{Acquire} = T^{-1}\sum_t a_{t,t}$ and trajectory-wide $\mathrm{AvgSeen} = T^{-1}\sum_t t^{-1}\sum_{i\le t} a_{t,i}$ supplement them. BWT uses each method's own task-arrival reference and is read alongside ACC and Acquire. Appendix B provides task lists, implementation settings, and complete cohort definitions.

5.2 BROAD COMPARISON ACROSS BACKBONES AND TASK HORIZONS

We first evaluate whether the fixed BiFilt-Sym rule remains effective across model scales and continual horizons. The broad benchmark retains the established $\mathcal{B}_{\mathrm{SVD}}$ first-stage convention of the existing continual-merging benchmark pipeline, together with ten matched task orders (seeds 42–51) for each of three CLIP backbones and three task-suite lengths. All methods within a setting use the same expert checkpoints, task orders, test sets, and evaluation procedure; methods with recovery use their stated configurations.

The role of $\mathcal{B}_{\mathrm{SVD}}$ here is protocol consistency rather than a claim that the zero-matrix singular vectors encode historical knowledge. We therefore do not compare the absolute values of this benchmark cohort with the audited $\mathcal{B}_0$ cohort to estimate an initialization effect; first-stage initialization is assessed only by the matched diagnostic in Appendix C.1.

Under this protocol, BiFilt-Sym obtains the highest final ACC among the evaluated merging methods in all nine backbone–suite settings. Relative to the right-sided filtering baseline with the same

Table 2: **Filtering structure under the audited $\mathcal{B}_0$ initialization.** TA8 results are mean $\pm$ sample SD across complete runs. ViT-B/32 uses five seeds (42, 43, 44, 46, 47), and ViT-B/16 uses three (43, 44, 46). All rows share historical rank 128 and rank-64, 50-step recovery. Bold marks the highest unrounded mean within each backbone and metric. A displayed SD of 0.0 may represent a nonzero value below 0.05.

Backbone	Filter	ACC (%) ↑	BWT (pp) ↑	AvgSeen (%) ↑	Acquire (%) ↑
ViT-B/32	RIGHT	$83.3_{\pm 0.3}$	$-2.3_{\pm 0.7}$	$88.6_{\pm 1.4}$	$85.3_{\pm 0.7}$
	LEFT	$83.6_{\pm 0.4}$	$-1.7_{\pm 0.5}$	$88.5_{\pm 1.2}$	$85.1_{\pm 0.7}$
	HARD	$83.2_{\pm 0.4}$	$\mathbf{-1.6}_{\pm 0.6}$	$88.2_{\pm 1.3}$	$84.6_{\pm 0.8}$
	BiFilt-Sym	$\mathbf{83.8}_{\pm 0.3}$	$-1.9_{\pm 0.5}$	$\mathbf{88.7}_{\pm 1.3}$	$\mathbf{85.4}_{\pm 0.7}$
ViT-B/16	RIGHT	$87.1_{\pm 0.2}$	$-1.4_{\pm 0.0}$	$91.0_{\pm 0.7}$	$\mathbf{88.4}_{\pm 0.2}$
	LEFT	$87.2_{\pm 0.3}$	$\mathbf{-1.0}_{\pm 0.1}$	$90.8_{\pm 0.7}$	$88.1_{\pm 0.2}$
	HARD	$86.7_{\pm 0.3}$	$-1.1_{\pm 0.2}$	$90.5_{\pm 0.7}$	$87.7_{\pm 0.2}$
	BiFilt-Sym	$\mathbf{87.3}_{\pm 0.2}$	$-1.2_{\pm 0.1}$	$\mathbf{91.0}_{\pm 0.7}$	$88.3_{\pm 0.2}$

recovery configuration, the paired mean ACC improvement ranges from 0.2 to 1.6 pp across the nine settings, while the BWT difference ranges from $+0.2$ to $+0.9$ pp. On ViT-B/32 TA8, the paired improvements are $+0.4$ pp in ACC and $+0.5$ pp in BWT, with positive ACC and BWT differences in all ten matched task orders. The ACC advantage also persists on the 14- and 20-task suites and across all three evaluated CLIP backbones, indicating that the benefit is not confined to a particular model scale or task horizon.

5.3 Filtering Structure Under the Audited Initialization

The broad benchmark retains the established first-stage convention for consistency with prior methods. For mechanistic analysis, however, the arbitrary singular directions produced by the SVD of a zero historical update constitute an unnecessary confounding factor. We therefore use the audited $\mathcal{B}_0$ initialization in the following structural controls: when the historical update has zero effective rank, no historical projector is constructed.

Table 2 compares right-sided, left-sided, hard bilateral, and symmetric filtering under an identical recovery budget. We identify RIGHT with $(\alpha, \beta) = (0, 1)$, LEFT with $(1, 0)$, HARD with $(0, 0)$, and BiFilt-Sym with $(0.5, 0.5)$.

Partial retention improves acquisition and final accuracy. On ViT-B/32, BiFilt-Sym achieves the highest ACC in all five evaluated seeds. Relative to HARD, it improves final ACC by 0.6 pp, Acquire by 0.8 pp, and AvgSeen by 0.6 pp, while BWT decreases by 0.3 pp. The same ACC pattern is reproduced on ViT-B/16, where BiFilt-Sym exceeds HARD by 0.5 pp. These results show that fully suppressing both cross blocks sacrifices useful plasticity on the standard eight-task sequences, whereas partial retention improves task acquisition and final performance.

The two cross directions are not interchangeable. LEFT and RIGHT differ only in which cross interaction is retained. On B/32, LEFT exceeds RIGHT by 0.3 pp ACC and 0.5 pp BWT; on B/16, the corresponding differences are 0.1 and 0.4 pp. The asymmetry becomes larger in the long-horizon diagnostic reported below. Together with the ordered-pair evidence in Section 3, these results support explicitly modeling the HC and CH interactions rather than collapsing them into a single cross term. They do not, however, establish a universal ordering for every task or layer.

5.4 Historical Core, Recovery, and Rank

For the historical-core control, we vary only the HH coefficient:

$$\mathcal{F}^{(w)}(\tau) = w\tau_{\parallel\parallel} + 0.5\tau_{\parallel\perp} + 0.5\tau_{\perp\parallel} + \tau_{\perp\perp}.$$

The default BiFilt-Sym filter has $w = 0$; positive w relaxes the selected-core preservation constraint in this ablation.

Table 3: **BiFilt design ablations under audited $\mathcal{B}_0$.** All controls use ViT-B/32 TA8 and seeds 43, 44, and 46. For historical-core retention, "Off" disables recovery and "On" uses 50-step, rank-64 recovery. Light recovery uses 10 steps and rank 16. The historical-rank sweep fixes BiFilt-Sym with 50-step, rank-64 recovery. Values are mean ± sample SD except the reported rank means.

Control	Setting	ACC (%) ↑	BWT (pp) ↑
Historical-core (HH) retention			
Recovery off	$w = 0$	$\mathbf{80.8}_{\pm 1.1}$	$\mathbf{-2.1}_{\pm 0.6}$
	$w = 0.25$	$79.6_{\pm 1.2}$	$-3.3_{\pm 0.7}$
	$w = 1/3$	$78.9_{\pm 1.2}$	$-3.9_{\pm 0.7}$
50 steps / rank 64	$w = 0$	$\mathbf{83.8}_{\pm 0.3}$	$\mathbf{-2.0}_{\pm 0.3}$
	$w = 0.25$	$83.6_{\pm 0.4}$	$-2.1_{\pm 0.2}$
	$w = 1/3$	$83.6_{\pm 0.4}$	$-2.1_{\pm 0.2}$
Light recovery: 10 steps / rank 16			
	RIGHT	$80.0_{\pm 1.1}$	$-2.6_{\pm 0.9}$
	LEFT	$81.8_{\pm 0.4}$	$-2.9_{\pm 0.6}$
	BiFilt-Sym	$\mathbf{81.9}_{\pm 0.8}$	$\mathbf{-1.8}_{\pm 0.5}$
Historical rank: BiFilt-Sym, 50 steps / rank 64			
	64	**84.0**	−2.5
	128	83.8	−2.0
	256	82.5	**−1.6**

Table 3 summarizes the $\mathcal{B}_0$ design controls on ViT-B/32 TA8 over seeds 43, 44, and 46. Without recovery, retaining HH with weight 0.25 instead of zero reduces ACC by 1.2 pp and BWT by 1.2 pp; weight $1/3$ is worse still (Table 6). Recovery attenuates but does not reverse this pattern. We therefore remove HH in the filtered proposal.

For BiFilt-Sym, 50-step rank-64 recovery raises ACC from 80.8% to 83.8% in this cohort, so it contributes substantially to the complete method. With only 10 steps and rank 16, BiFilt-Sym retains the best ACC and BWT among RIGHT, LEFT, and BiFilt-Sym; the filtering comparison is therefore not explained solely by a strong recovery budget (Table 7). In the revised historical-rank control, rank 64 gives the highest ACC (84.0%), while rank 256 gives the best BWT (-1.6 pp) at lower ACC (82.5%). Rank 128 is the fixed compromise used elsewhere.

Long-horizon stress test. On one ViT-B/32 TALL20 order, HARD provides stronger final retention (72.4% ACC, -8.1 pp BWT) than BiFilt-Sym (71.8%, -9.2 pp), whereas BiFilt-Sym retains slightly higher AvgSeen and Acquire. LEFT also exceeds RIGHT in both ACC and BWT. This diagnostic suggests that stricter cross-block filtering may become more favorable for final retention over longer horizons, but the single order does not establish a general horizon-dependent trend. Full results are given in Table 5.

6 Conclusion and Limitations

BiFilt distinguishes the two directional cross interactions while preserving the selected historical input–output core in its pre-recovery proposal. Ordered-pair experiments show that the complementary-output/historical-input block can carry useful new-task information. With fixed cross-block coefficients, BiFilt-Sym has the highest final accuracy among evaluated merging methods in all nine settings of the $\mathcal{B}_{\mathrm{SVD}}$ broad benchmark; its advantage over hard bilateral filtering is also reproduced on two eight-task backbones under $\mathcal{B}_0$. In one 20-task order, hard filtering has stronger final retention, while partial retention improves acquisition and average-seen performance. Task-level effects remain heterogeneous, including a repeatable degradation on SVHN in the broad benchmark. Recovery can change the final block geometry, so exact core preservation applies to the filtered proposal. These findings characterize both the value and observed limits of controlled cross-block retention in data-free continual model merging.

AI USE STATEMENT

Generative AI tools assisted with developing and checking mathematical formulations and derivations, refining hypotheses and experimental methodology, interpreting reported aggregate results, and synthesizing related work. They also assisted with language and LaTeX editing, manuscript organization, restructuring tables, creating or revising scientific figures and plots from supplied measurements, and checking submission requirements. AI assistance helped organize ablation controls and compute summaries from supplied records; it did not run the underlying model-merging experiments or supply unmeasured values. The authors reviewed the AI-assisted text, mathematics, citations, tables, figures, and computed summaries against the manuscript sources and experimental records and take responsibility for the final content.

ETHICS STATEMENT

The described study concerns merging existing model checkpoints for established visual-classification benchmarks and does not introduce a new human-subject study. Merged models can inherit biases, errors, and harmful capabilities from their source models. The absence of task data during merging is an access constraint, not a privacy guarantee, and parameter sharing may still expose information encoded in source checkpoints. Deployment therefore requires application-specific evaluation beyond aggregate benchmark accuracy. Reuse and redistribution of checkpoints and datasets must respect their original licenses and access conditions. Taskwise evaluation accompanies aggregate benchmark results.

REPRODUCIBILITY STATEMENT

Section 4 and Algorithm 1 define the complete bilateral filtering, recovery, and continual-merging procedure while treating the first-stage rule as an explicit protocol choice.

The broad benchmark retains the established $\mathcal{B}_{\mathrm{SVD}}$ implementation convention, whereas the main structural and mechanistic analyses use the audited rank-zero initialization $\mathcal{B}_0$. These archived cohorts are reported separately and are not compared through their absolute accuracies to estimate initialization effects. A dedicated matched diagnostic in Appendix C.1 compares rank-zero initialization against three independently generated first-stage orthogonal bases while holding the subsequent BiFilt path and stage-wise random-number protocol fixed. Supplementary analyses derived from the archived benchmark cohort retain $\mathcal{B}_{\mathrm{SVD}}$ and are identified accordingly.

REFERENCES

Runxi Cheng, Feng Xiong, Yongxian Wei, Wanyun Zhu, and Chun Yuan. Whoever started the interference should end it: Guiding data-free model merging via task vectors. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10121–10143, 2025.

Donato Crisostomi, Alessandro Zirilli, Antonio Andrea Gargiulo, Maria Sofia Bucarelli, Simone Scardapane, Fabrizio Silvestri, Iacopo Masi, and Emanuele Rodolà. MASS: MoErging through adaptive subspace selection. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/21f1dd8ee699b699f95b589ddf78fa0c-Abstract-Conference.html.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3762–3773, 2020.

Reinhard Heckel. Provable continual learning via sketched Jacobian approximations. In *Proceedings of the Twenty-Fifth International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 10448–10470, 2022.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023.

Chanhyuk Lee, Jiho Choi, Chanryeol Lee, Donggyun Kim, and Seunghoon Hong. AdaRank: Adaptive rank pruning for enhanced model merging. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0b77a07c3c12b52121b3f42f96875726-Abstract-Conference.html.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, pp. 6467–6476, 2017.

Daniel Marczak, Bartlomiej Twardowski, Tomasz Trzcinski, and Sebastian Cygert. MagMax: Leveraging model merging for seamless continual learning. In *European Conference on Computer Vision*, pp. 379–395. Springer, 2024. doi: 10.1007/978-3-031-73013-9_22.

Daniel Marczak, Simone Magistri, Sebastian Cygert, Bartłomiej Twardowski, Andrew D. Bagdanov, and Joost Van De Weijer. No task left behind: Isotropic model merging with common and task-specific subspaces. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43177–43199, 2025.

Michael Matena and Colin Raffel. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Zihuan Qiu, Zhiyang Liao, Chiyuan He, Yi Xu, Fanman Meng, Linfeng Xu, Qingbo Wu, and Hongliang Li. SAFE-Merge: Data-free continual model merging with general knowledge preservation, 2026a. URL https://arxiv.org/abs/2608.01184.

Zihuan Qiu, Lei Wang, Yang Cao, Runtong Zhang, Bing Su, Yi Xu, Fanman Meng, Linfeng Xu, Qingbo Wu, and Hongliang Li. Null-space filtering for data-free continual model merging: Preserving stability, promoting plasticity. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/d0da30e312b75a3fffd9e9191f8bc1b0-Abstract-Conference.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=3AOj0RCNC2.

George Stoica, Pratik Ramesh, Boglarka Ecsedi, Leshem Choshen, and Judy Hoffman. Model merging with SVD to tie the knots. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/0d4f8a5109c5083b5307fcd0bddae7a7-Abstract-Conference.html.

Derek Tam, Mohit Bansal, and Colin Raffel. Merging by matching models in task parameter subspaces. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=qNGo6ghWFB.

Anke Tang, Enneng Yang, Li Shen, Yong Luo, Han Hu, Lefei Zhang, Bo Du, and Dacheng Tao. Merging on the fly without retraining: A sequential approach to scalable continual model merging. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Anke Tang, Li Shen, Yong Luo, Shuai Xie, Han Hu, Lefei Zhang, Bo Du, and Dacheng Tao. Zero-shot sparse mixture of low-rank experts construction from pre-trained foundation models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):1145–1157, 2026. doi: 10.1109/TPAMI.2025.3612480.

Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. Training networks in null space of feature covariance for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 184–193, 2021.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998, 2022.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A. Raffel, and Mohit Bansal. TIES-Merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. AdaMerging: Adaptive model merging for multi-task learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=nZP6NgD3QY.

Enneng Yang, Anke Tang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, and Jie Zhang. Continual model merging without data: Dual projections for balancing stability and plasticity. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

APPENDIX

A ADDITIONAL DERIVATIONS

Throughout this appendix, all projectors are fixed and orthogonal. For matrices, $\|\cdot\|_2$ denotes the spectral norm, $\|\cdot\|_F$ the Frobenius norm, and $\langle X, Y\rangle_F = \mathrm{tr}(X^\top Y)$ the Frobenius inner product. The results establish the preservation objective and projection identities, then relate parameter geometry to fixed-layer responses under explicit alignment conditions. Layer superscripts are omitted in fixed-layer calculations unless needed for clarity.

A.1 FROM TASK LOSS TO BILATERAL GEOMETRY

Jacobian- and subspace-based continual-learning methods motivate viewing parameter directions that leave historical behavior unchanged to first order as locally non-interfering (Farajtabar et al., 2020; Heckel, 2022; Saha et al., 2021; Wang et al., 2021). We first define this notion from task behavior, independently of the parameter-derived singular spaces used by BiFilt.

Functional first-order safety. Fix a reference network and a matrix layer ℓ. For a protected collection of historical examples, let $h_\ell^{(n)}$ denote the input to layer ℓ and $g_\ell^{(n)}$ the gradient of the corresponding loss with respect to the layer output. Define the historical feature and output-gradient subspaces by

$$\mathcal{H}_\ell := \mathrm{span}\{h_\ell^{(n)}\}_n, \qquad \mathcal{G}_\ell := \mathrm{span}\{g_\ell^{(n)}\}_n. \tag{12}$$

Let $P_{\mathcal{H}_\ell}$ and $P_{\mathcal{G}_\ell}$ denote their orthogonal projectors.

We define the *span-wise first-order safe update space* as

$$\mathfrak{S}_\ell := \left\{ Z_\ell : g^\top Z_\ell h = 0,\ \forall h \in \mathcal{H}_\ell,\ \forall g \in \mathcal{G}_\ell \right\}. \tag{13}$$

This span-wise condition is stronger than requiring $(g_\ell^{(n)})^\top Z_\ell h_\ell^{(n)} = 0$ only for matched examples, but it admits a clean subspace characterization.

Proof of Theorem 1. For any $h \in \mathcal{H}_\ell$ and $g \in \mathcal{G}_\ell$,

$$h = P_{\mathcal{H}_\ell} h, \qquad g = P_{\mathcal{G}_\ell} g.$$

Hence

$$g^\top Z_\ell h = g^\top P_{\mathcal{G}_\ell} Z_\ell P_{\mathcal{H}_\ell} h.$$

Therefore

$$P_{\mathcal{G}_\ell} Z_\ell P_{\mathcal{H}_\ell} = 0$$

implies $g^\top Z_\ell h = 0$ for every $h \in \mathcal{H}_\ell$ and $g \in \mathcal{G}_\ell$.

Conversely, suppose $g^\top Z_\ell h = 0$ for all such h and g. For every $h \in \mathcal{H}_\ell$, the vector $P_{\mathcal{G}_\ell} Z_\ell h$ has zero inner product with every vector in $\mathcal{G}_\ell$ while itself belonging to $\mathcal{G}_\ell$. It must therefore be zero. Thus

$$P_{\mathcal{G}_\ell} Z_\ell P_{\mathcal{H}_\ell} = 0.$$

Finally, $P_{\mathcal{G}_\ell} Z_\ell P_{\mathcal{H}_\ell} = 0$ is equivalent to

$$Z_\ell \mathcal{H}_\ell \subseteq \mathcal{G}_\ell^\perp,$$

which completes the characterization. □

A parameter-only bilateral proxy. The functional spaces $\mathcal{H}_\ell$ and $\mathcal{G}_\ell$ depend on historical examples and are therefore unavailable in DFCMM. Let U_ℓ and V_ℓ instead denote the retained left and right singular bases of the accumulated historical parameter update, and define the bilateral feasible set

$$\mathcal{A}_\ell := \left\{ Z_\ell : U_\ell^\top Z_\ell V_\ell = 0 \right\}. \tag{14}$$

Corollary 1 (Bilateral constraint as a sufficient safety condition)**.** *Suppose the parameter-derived subspaces cover the historical functional subspaces:*

$$\mathcal{H}_\ell \subseteq \mathrm{span}(V_\ell), \qquad \mathcal{G}_\ell \subseteq \mathrm{span}(U_\ell). \tag{15}$$

Then

$$\mathcal{A}_\ell \subseteq \mathfrak{S}_\ell.$$

Proof. For any $Z_\ell \in \mathcal{A}_\ell$, $h \in \mathcal{H}_\ell$, and $g \in \mathcal{G}_\ell$, the coverage assumptions imply that $h = V_\ell a$ and $g = U_\ell b$ for some a, b. Therefore

$$g^\top Z_\ell h = b^\top U_\ell^\top Z_\ell V_\ell a = 0.$$

Hence $Z_\ell \in \mathfrak{S}_\ell$. □

Corollary 1 should be read as a conditional link between functional and parameter geometry. The coverage condition is not implied by the parameter SVD and cannot be verified from model parameters alone.

Downstream enlargement of functional safety. The local space $\mathfrak{S}_\ell$ need not exhaust all first-order safe responses at layer ℓ. Consider, for clarity, a linear reference chain with weights $W_1, \ldots, W_L$. For $j \geq \ell$, define the propagation operator

$$T_{\ell,\ell} = I, \qquad T_{j,\ell} = W_j W_{j-1} \cdots W_{\ell+1}, \quad j > \ell.$$

For depth m, define the downstream-safe response-direction space

$$\mathcal{K}_\ell^{(m)} := \sum_{j=\ell}^{\ell+m-1} T_{j,\ell}^{-1}\left[\mathcal{G}_j^\perp\right], \tag{16}$$

where

$$T^{-1}[E] := \{v : Tv \in E\}$$

denotes a set preimage, not a matrix inverse or pseudoinverse.

The corresponding update space is

$$\mathfrak{S}_\ell^{(m)} := \left\{ Z_\ell : Z_\ell \mathcal{H}_\ell \subseteq \mathcal{K}_\ell^{(m)} \right\}. \tag{17}$$

Proposition 1 (Downstream enlargement of the safe space)**.** *For the linear reference chain, every $Z_\ell \in \mathfrak{S}_\ell^{(m)}$ has zero first-order effect on the protected historical losses. Moreover,*

$$\mathfrak{S}_\ell^{(1)} \subseteq \mathfrak{S}_\ell^{(2)} \subseteq \cdots \subseteq \mathfrak{S}_\ell^{(L-\ell+1)}, \tag{18}$$

and

$$\mathfrak{S}_\ell^{(1)} = \mathfrak{S}_\ell.$$

Proof. For a protected example and any $j \geq \ell$, backpropagation through the linear reference chain gives

$$g_\ell = T_{j,\ell}^\top g_j.$$

If

$$\delta \in T_{j,\ell}^{-1}\left[\mathcal{G}_j^\perp\right],$$

then $T_{j,\ell}\delta \in \mathcal{G}_j^\perp$ and therefore

$$g_\ell^\top \delta = g_j^\top T_{j,\ell}\delta = 0.$$

Each summand in $\mathcal{K}_\ell^{(m)}$ is thus contained in the kernel of the historical first-order loss differential at layer ℓ. Because this kernel is a linear subspace, their sum $\mathcal{K}_\ell^{(m)}$ is also safe.

If $Z_\ell \in \mathfrak{S}_\ell^{(m)}$, then for every $h \in \mathcal{H}_\ell$,

$$Z_\ell h \in \mathcal{K}_\ell^{(m)},$$

so $g_\ell^\top Z_\ell h = 0$ for every protected example.

Increasing m adds another preimage subspace to $\mathcal{K}_\ell^{(m)}$, which proves the nesting in Eq. 18. For $m = 1$, $T_{\ell,\ell} = I$ and hence

$$\mathcal{K}_\ell^{(1)} = \mathcal{G}_\ell^\perp.$$

Equation 17 therefore reduces to the safe-space characterization in Theorem 1. □

This result formalizes the fact that a perturbation need not be locally orthogonal to the current-layer historical gradient space to be safe: it may instead propagate into a direction that becomes orthogonal to a deeper historical gradient space. The inclusions in Eq. 18 are strict only when the additional preimages contribute new directions.

Parameter-derived approximation of downstream safety. The functional gradient spaces in Eq. 16 are unavailable at merge time. Under the coverage condition

$$\mathcal{G}_j \subseteq \mathrm{span}(U_j),$$

we have

$$\mathrm{span}(U_j)^\perp \subseteq \mathcal{G}_j^\perp.$$

Hence the parameter-derived response space

$$\widehat{\mathcal{K}}_\ell^{(m)} := \sum_{j=\ell}^{\ell+m-1} T_{j,\ell}^{-1}\big[\mathrm{span}(U_j)^\perp\big] \tag{19}$$

satisfies

$$\widehat{\mathcal{K}}_\ell^{(m)} \subseteq \mathcal{K}_\ell^{(m)}.$$

If additionally $\mathcal{H}_\ell \subseteq \mathrm{span}(V_\ell)$, then

$$\widehat{\mathcal{A}}_\ell^{(m)} := \Big\{ Z_\ell : \mathrm{range}(Z_\ell V_\ell) \subseteq \widehat{\mathcal{K}}_\ell^{(m)} \Big\} \tag{20}$$

is a sufficient parameter-derived subset of $\mathfrak{S}_\ell^{(m)}$.

For $m = 1$,

$$\widehat{\mathcal{K}}_\ell^{(1)} = \mathrm{span}(U_\ell)^\perp,$$

and $\widehat{\mathcal{A}}_\ell^{(1)}$ reduces exactly to the bilateral feasible set $\mathcal{A}_\ell$ in Eq. 14. Thus the local bilateral constraint can be viewed as the first, parameter-computable member of a broader hierarchy of downstream-safe directions.

Projection onto the bilateral feasible set. The bilateral constraint specifies which updates are admissible but does not yet determine which admissible update should be selected.

Corollary 2 (Projection onto the bilateral feasible set)**.** *Let τ_ℓ be an incoming layer update. The unique solution of*

$$\min_{Z_\ell \in \mathcal{A}_\ell} \|Z_\ell - \tau_\ell\|_F^2$$

is

$$Z_\ell^\star = \tau_\ell - P_{U_\ell} \tau_\ell P_{V_\ell}. \tag{21}$$

Equivalently,

$$Z_\ell^\star = \tau_{\parallel\perp}^{(\ell)} + \tau_{\perp\parallel}^{(\ell)} + \tau_{\perp\perp}^{(\ell)} = \mathcal{F}_{1,1}(\tau_\ell).$$

Proof. Define

$$\Pi_\ell(Z) := P_{U_\ell} Z P_{V_\ell}.$$

Because P_{U_ℓ} and P_{V_ℓ} are orthogonal projectors, Π_ℓ is self-adjoint and idempotent under the Frobenius inner product, and is therefore an orthogonal projector on the matrix space. Moreover,

$$\ker(\Pi_\ell) = \mathcal{A}_\ell.$$

Hence $I - \Pi_\ell$ is the orthogonal projector onto $\mathcal{A}_\ell$, yielding

$$Z_\ell^\star = (I - \Pi_\ell)(\tau_\ell) = \tau_\ell - P_{U_\ell} \tau_\ell P_{V_\ell}.$$

□

Corollary 2 shows that Euclidean projection onto the bilateral feasible set removes only the jointly historical block. The regularized objective in Eq. 7 goes further: while remaining in the same feasible set, it separately penalizes the two cross interactions, yielding the BiFilt coefficients α and β.

A fixed-layer loss expansion. The filtering criterion preserves a selected parameter mapping. We now relate its geometry to local task loss under explicit smoothness and alignment conditions. The conceptual task objective in Section 4.1 averages risk over the seen distributions. Write $\overline{\mathcal{D}}_t = t^{-1}\sum_{i=1}^{t}\mathcal{D}_i$ for this mixture. Fix a reference network, one matrix layer ℓ, and all other parameters. For a sample (x, y), let $h^{(\ell)}(x)$ be the input to that layer and let $z^{(\ell)}(x)$ be its output before downstream operations. With a fixed bias, changing its weight by $\Delta^{(\ell)}$ changes that output exactly by $\Delta^{(\ell)}h^{(\ell)}(x)$. Let $\varphi^{(\ell)}_{x,y}(z)$ denote the loss of the downstream computation as a function of this output, with other inputs held fixed. Define $g^{(\ell)} = \nabla_z\varphi^{(\ell)}_{x,y}$ and $H^{(\ell)} = \nabla^2_z\varphi^{(\ell)}_{x,y}$ at the reference output.

Assuming twice continuous differentiability locally and an integrable bound that permits averaging the Taylor remainder, the risk change is

$$\delta\mathcal{R}_t(\Delta^{(\ell)}) = \mathbb{E}_{\overline{\mathcal{D}}_t}\left[(g^{(\ell)})^\top\Delta^{(\ell)}h^{(\ell)} + \frac{1}{2}(\Delta^{(\ell)}h^{(\ell)})^\top H^{(\ell)}(\Delta^{(\ell)}h^{(\ell)})\right] + o(\|\Delta^{(\ell)}\|_F^2). \tag{22}$$

Suppressing the layer index and the remainder gives

$$\delta\mathcal{R}_t(\Delta) \approx \mathbb{E}\left[g^\top\Delta h + \tfrac{1}{2}(\Delta h)^\top H(\Delta h)\right], \tag{23}$$

The expectation is over the seen-task mixture. The gradient term retains the sign of the response's alignment with the loss gradient. The curvature term weights output directions through $H^{(\ell)}$; this Hessian need not be positive semidefinite. For column-wise vectorization, the sample quadratic term also satisfies

$$(\Delta h)^\top H(\Delta h) = \mathrm{vec}(\Delta)^\top\left(hh^\top \otimes H\right)\mathrm{vec}(\Delta). \tag{24}$$

Let

$$\mathcal{I}_{\mathrm{blk}} = \{\|\|, \|\perp, \perp\|, \perp\perp\},$$

and define

$$\Delta_w = \sum_{b\in\mathcal{I}_{\mathrm{blk}}} w_b\tau_b.$$

Expanding the weighted update gives

$$\delta\mathcal{R}_t(\Delta_w) \approx \sum_{b\in\mathcal{I}_{\mathrm{blk}}} w_b\,\mathbb{E}[g^\top\tau_b h] + \frac{1}{2}\sum_{b,c\in\mathcal{I}_{\mathrm{blk}}} w_b w_c\,\mathbb{E}[(\tau_b h)^\top H(\tau_c h)]. \tag{25}$$

The $b \neq c$ terms generally remain: Frobenius orthogonality of the parameter blocks does not imply orthogonality of their activation responses under the loss Hessian.

Thus local loss geometry involves both input activation directions and output-side derivatives. These quantities depend on the task and reference network. The expansion considers a single-layer perturbation; simultaneous changes to several layers additionally produce cross-layer terms. For a weight shared across token positions, stack the output changes into $d = \mathrm{vec}(\Delta X)$, where the columns of X are token inputs, and use $g^\top d + \frac{1}{2}d^\top Hd$ with the stacked gradient and Hessian. This retains interactions between token positions.

Distinctions coupled by right-sided filtering. Consider the historical update $S_{t-1} = \mathrm{diag}(1, 0)$ with rank-one bases $U = V = e_1$. Let the layer input be $h = e_1$ and compare increments $\Delta_{\|\|} = se_1e_1^\top$ and $\Delta_{\perp\|} = se_2e_1^\top$. Both act on the same input direction and both are removed by right-sided filtering: $\Delta_{\|\|}P_V^\perp = \Delta_{\perp\|}P_V^\perp = 0$. Their output responses are se_1 and se_2, respectively. For the illustrative local loss

$$\varphi(z) = \tfrac{1}{2}(z_1 - y_1)^2 + \tfrac{\kappa}{2}(z_2 - y_2)^2, \qquad \kappa > 0, \tag{26}$$

at a reference output $z = y$, their exact loss changes are $s^2/2$ and $\kappa s^2/2$. For $0 < \kappa < 1$, the $\perp\|$ response is less costly; for $\kappa > 1$, it is more costly. The distinction comes from output sensitivity despite identical input alignment. With $h = e_2$, the same comparison applies to $\Delta_{\|\perp} = se_1e_2^\top$ and $\Delta_{\perp\perp} = se_2e_2^\top$, both of which the right-sided filter retains.

From task geometry to a parameter-only prior. The task derivatives in the expansion depend on activations and the downstream network. Historical singular spaces summarize accumulated parameter changes. Connecting the two requires alignment conditions; the filter itself is obtained from the explicit preservation criterion in Eq. 7. The following bound states one such conditional connection.

A conditional bound on first-order loss change. Fix a differentiable downstream sample loss φ at output z, with $g = \nabla\varphi(z)$ and layer input h. Write $F = \mathcal{F}_{\alpha,\beta}(\tau)$ for this proof. The directional derivative is exactly $\frac{d}{ds}\varphi(z + sFh)\big|_{s=0} = g^\top Fh$. Suppose

$$\|P_V^\perp h\|_2 \le \epsilon_h \|h\|_2, \qquad \|P_U^\perp g\|_2 \le \epsilon_g \|g\|_2.$$

Using the left and right support of each block gives

$$\begin{aligned} g^\top Fh = \beta (P_U g)^\top \tau_{\parallel\perp}(P_V^\perp h) + \alpha (P_U^\perp g)^\top \tau_{\perp\parallel}(P_V h) \\ + (P_U^\perp g)^\top \tau_{\perp\perp}(P_V^\perp h). \end{aligned} \tag{27}$$

Apply $|a^\top Mb| \le \|a\|_2 \|M\|_2 \|b\|_2$ to each term and use contractivity of orthogonal projectors to obtain

$$|g^\top Fh| \le \|g\|_2 \|h\|_2 \big(\beta\epsilon_h \|\tau_{\parallel\perp}\|_2 + \alpha\epsilon_g \|\tau_{\perp\parallel}\|_2 + \epsilon_h\epsilon_g \|\tau_{\perp\perp}\|_2\big). \tag{28}$$

The two cross blocks carry an input or output residual factor, respectively; the doubly complementary block carries both. These are assumptions on the sample's activation and gradient, not consequences of parameter SVD. The bound controls magnitude, not the sign of the loss change.

If $\nabla\varphi$ is L-Lipschitz along the segment from z to $z + sFh$, the fundamental theorem of calculus and the Lipschitz bound imply

$$\big|\varphi(z + sFh) - \varphi(z) - s\, g^\top Fh\big| \le \frac{Ls^2}{2} \|Fh\|_2^2. \tag{29}$$

Indeed, integrate $s[\nabla\varphi(z + usFh) - g]^\top Fh$ over $u \in [0, 1]$. This separates first-order sensitivity from the finite-step remainder. Both statements hold at a fixed layer with the remaining network fixed.

A.2 Orthogonality of the Bilateral Decomposition

Commuting orthogonal projections. For matrices X, Z, $\langle X, P_U Z\rangle_F = \langle P_U X, Z\rangle_F$ and $\langle X, ZP_V\rangle_F = \langle XP_V, Z\rangle_F$, since both projectors are symmetric. Their idempotence follows from $P_U^2 = P_U$ and $P_V^2 = P_V$. Left and right projection commute as operators on the matrix space because

$$P_U(ZP_V) = (P_U Z)P_V.$$

Hence the four combinations of the two projections and their complements form mutually orthogonal projection operators. They give the unique joint orthogonal decomposition for fixed P_U, P_V. This claim fixes the selected subspaces; an SVD cutoff through a tied singular-value group requires a choice of subspace.

Let

$$L_\parallel = P_U, \qquad L_\perp = P_U^\perp, \qquad R_\parallel = P_V, \qquad R_\perp = P_V^\perp.$$

Since $P_U + P_U^\perp = I_{d_{\text{out}}}$ and $P_V + P_V^\perp = I_{d_{\text{in}}}$,

$$\tau = \sum_{a,b\in\{\parallel,\perp\}} L_a \tau R_b. \tag{30}$$

For two blocks $\tau_{ab} = L_a\tau R_b$ and $\tau_{cd} = L_c\tau R_d$,

$$\langle \tau_{ab}, \tau_{cd}\rangle_F = \operatorname{tr}\left(R_b \tau^\top L_a L_c \tau R_d\right). \tag{31}$$

If $a \ne c$, orthogonality of the left projectors gives $L_a L_c = 0$. If $a = c$ but $b \ne d$, cyclic invariance of the trace yields a factor $R_d R_b = 0$. Hence any two distinct blocks are orthogonal under the Frobenius inner product, establishing Eq. 32.

$$\|\tau\|_F^2 = \|\tau_{\parallel\parallel}\|_F^2 + \|\tau_{\parallel\perp}\|_F^2 + \|\tau_{\perp\parallel}\|_F^2 + \|\tau_{\perp\perp}\|_F^2. \tag{32}$$

This argument depends only on the projector structure and does not require the accumulated update S_{t-1} itself to have rank k.

A characterization under coordinate equivariance. Fix P_U, P_V. Let a linear map $\mathcal{A}$ on the matrix space satisfy $\mathcal{A}(OZQ^\top) = O\mathcal{A}(Z)Q^\top$ for every pair of orthogonal matrices O, Q that independently preserve the selected output/input subspaces and their complements. Then $\mathcal{A}$ multiplies each nonempty interaction block by a scalar.

To prove this, choose orthonormal coordinates adapted to the subspaces. For a matrix unit E_{ij}, independent sign flips of output and input coordinates transform it with character $s_i t_j$. Equivariance forces $\mathcal{A}(E_{ij})$ to have the same character. Since the signs are independent, every entry except (i, j) vanishes, giving $\mathcal{A}(E_{ij}) = c_{ij}E_{ij}$. Permuting coordinates within an output group or within an input group makes c_{ij} constant within each of the four blocks. Linearity completes the proof. Conversely, blockwise scalar maps have the stated equivariance because these rotations commute with the projectors. Empty blocks contribute zero. This result characterizes filters under the stated linearity and equivariance assumptions; it does not select the coefficients.

A.3 Operator Relations and Exact Norm Identities

The boundary cases are

$$\begin{aligned} \mathcal{F}_{0,1}(\tau) &= \tau P_V^\perp, && \text{right-sided hard filtering,} \\ \mathcal{F}_{1,0}(\tau) &= P_U^\perp \tau, && \text{left-sided hard filtering,} \\ \mathcal{F}_{0,0}(\tau) &= P_U^\perp \tau P_V^\perp, && \text{two-sided hard filtering,} \\ \mathcal{F}_{1,1}(\tau) &= \tau - \tau_{\|\|}, && \|\| \text{ removal only.} \end{aligned} \tag{33}$$

The equivalent projector form is

$$\mathcal{F}_{\alpha,\beta}(\tau) = \alpha P_U^\perp \tau + \beta \tau P_V^\perp + (1 - \alpha - \beta) P_U^\perp \tau P_V^\perp. \tag{34}$$

Using

$$P_U^\perp \tau = \tau_{\perp\|} + \tau_{\perp\perp}, \qquad \tau P_V^\perp = \tau_{\|\perp} + \tau_{\perp\perp}, \qquad P_U^\perp \tau P_V^\perp = \tau_{\perp\perp},$$

and collecting block coefficients directly yields Eq. 34. In particular, the symmetric default in Section 4.3 is *exactly* the arithmetic mean of the left- and right-sided hard filters rather than an approximation.

Relative to right-sided hard filtering, the symmetric default reallocates the cross blocks as

$$\mathcal{F}_{0.5,0.5}(\tau) - \tau P_V^\perp = \tfrac{1}{2}\tau_{\perp\|} - \tfrac{1}{2}\tau_{\|\perp}. \tag{35}$$

Relation to products of soft projectors. A natural alternative is to apply soft projectors independently on the left and right. Expanding such a product gives

$$\begin{aligned} (P_U^\perp + \beta P_U)\tau(P_V^\perp + \alpha P_V) &= \tau_{\perp\perp} + \beta\tau_{\|\perp} + \alpha\tau_{\perp\|} + \alpha\beta\tau_{\|\|}, \\ \mathcal{F}_{\alpha,\beta}(\tau) &= (P_U^\perp + \beta P_U)\tau(P_V^\perp + \alpha P_V) - \alpha\beta\tau_{\|\|}. \end{aligned} \tag{36}$$

Thus, the proposed operator is not simply the product of two soft projectors: explicit removal of the $\|\|$ block is an additional component of the retention prior.

Likewise, a purely right-sided soft filter satisfies

$$\tau(P_V^\perp + \alpha P_V) = \tau_{\perp\perp} + \tau_{\|\perp} + \alpha\tau_{\perp\|} + \alpha\tau_{\|\|}, \tag{37}$$

whereas $\mathcal{F}_{\alpha,1}(\tau)$ omits the final $\|\|$ term. We therefore refer to $(0.5, 1)$ as $\perp\|$*-relaxed filtering* rather than single-sided soft filtering.

Energy decomposition and non-expansiveness. The four block projections are mutually orthogonal linear operators in the matrix Hilbert space. For fixed bases, $\mathcal{F}_{\alpha,\beta}$ scales the $(\|\|, \|\perp, \perp\|, \perp\perp)$ blocks by $(0, \beta, \alpha, 1)$, respectively. Consequently,

$$\begin{aligned} \|\mathcal{F}(\tau)\|_F^2 &= \|\tau_{\perp\perp}\|_F^2 + \beta^2\|\tau_{\|\perp}\|_F^2 + \alpha^2\|\tau_{\perp\|}\|_F^2, \\ \|\tau - \mathcal{F}(\tau)\|_F^2 &= \|\tau_{\|\|}\|_F^2 + (1-\beta)^2\|\tau_{\|\perp}\|_F^2 + (1-\alpha)^2\|\tau_{\perp\|}\|_F^2. \end{aligned} \tag{38}$$

For any two updates τ_1 and τ_2 evaluated using the same bases and coefficients,

$$\begin{aligned}\|\mathcal{F}(\tau_1) - \mathcal{F}(\tau_2)\|_F &= \|\mathcal{F}(\tau_1 - \tau_2)\|_F \\ &\leq \|\tau_1 - \tau_2\|_F.\end{aligned} \tag{39}$$

Thus, for fixed projectors, the filtering operator is non-expansive in Frobenius norm. In the continual procedure, the historical bases and recovery solution evolve with the accumulated model, so the fixed-projector result characterizes each filtering step.

Relation to two-sided projection penalties. For an arbitrary residual matrix Z, an unweighted two-sided penalty decomposes as

$$\|P_U Z\|_F^2 + \|Z P_V\|_F^2 = 2\|Z_{\|\|}\|_F^2 + \|Z_{\|\perp}\|_F^2 + \|Z_{\perp\|}\|_F^2. \tag{40}$$

This identity shows that two-sided projection penalties can already distinguish joint historical overlap from cross-subspace overlap. DOP additionally uses singular-value weighting and separate historical and incoming targets (Yang et al., 2025). BiFilt constrains the compressed historical map, then derives the retained update by balancing fidelity and two response penalties.

A.4 Conditional Responses and Limits of the Historical Proxy

The projector identities imply

$$\tau_{\|\perp} x = \tau_{\|\perp} P_V^{\perp} x, \qquad \tau_{\perp\|} x = \tau_{\perp\|} P_V x, \qquad \tau_{\perp\perp} x = \tau_{\perp\perp} P_V^{\perp} x. \tag{41}$$

For a perfectly history-aligned input $x = P_V x$, the first and third terms vanish. Multiplying the filter by P_V on the right or by P_U on the left therefore gives Eq. 42.

$$\mathcal{F}_{\alpha,\beta}(\tau) P_V = \alpha \tau_{\perp\|}, \qquad P_U \mathcal{F}_{\alpha,\beta}(\tau) = \beta \tau_{\|\perp}. \tag{42}$$

More generally, suppose a nonzero input satisfies

$$\|P_V^{\perp} x\|_2 \leq \epsilon \|x\|_2.$$

Then

$$\begin{aligned}\|\mathcal{F}(\tau)x\|_2 &\leq \alpha \|\tau_{\perp\|}\|_2 \|P_V x\|_2 + \left(\beta \|\tau_{\|\perp}\|_2 + \|\tau_{\perp\perp}\|_2\right) \|P_V^{\perp} x\|_2 \\ &\leq \left[\alpha \|\tau_{\perp\|}\|_2 + \epsilon \left(\beta \|\tau_{\|\perp}\|_2 + \|\tau_{\perp\perp}\|_2\right)\right] \|x\|_2,\end{aligned} \tag{43}$$

which establishes the conditional fixed-layer response bound.

The bound applies to a fixed layer and input, with relative residual ϵ measuring input alignment with $\mathrm{span}(V)$. The parameter-derived basis and the activation alignment are separate quantities. Applying the bound across a network would additionally require tracking how earlier-layer updates change each downstream input.

A.4.1 Historical responses to complementary updates

Consider

$$S_{t-1} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad \tau = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \qquad U = V = e_1. \tag{44}$$

Here $\tau = \tau_{\perp\perp}$. Nevertheless, for the input $x = e_2$,

$$\tau x = e_2 \neq 0. \tag{45}$$

Thus, this complementary parameter update produces a nonzero response on the historical sample.

Similarly, a block $\tau_{\|\|} = UCV^{\top}$ may either align with or oppose the accumulated update depending on the coefficient matrix C. The coefficient matrix determines sign agreement within the $\|\|$ block.

Finally, if

$$S_{t-1} = U \Sigma V^{\top} + E, \tag{46}$$

then

$$S_{t-1} P_V^{\perp} = E P_V^{\perp}, \quad P_U^{\perp} S_{t-1} = P_U^{\perp} E. \tag{47}$$

These residuals vanish when the selected bases span the full nonzero singular subspaces. The activation component outside $\mathrm{span}(V)$ remains a separate quantity, even at full rank. Historical singular subspaces therefore serve as parameter-space proxies for analyzing update alignment.

A.5 Unique Solution of the Preservation Objective

The feasible set in Eq. 7 is the linear subspace $P_U Z P_V = 0$. Since U, V have orthonormal columns, this is equivalent to $U^\top Z V = 0$: multiply by $U^\top, V$ in one direction and by $U, V^\top$ in the other. The norm identities $\|U^\top Z\|_F = \|P_U Z\|_F$ and $\|ZV\|_F = \|ZP_V\|_F$ follow by expanding squared norms as traces. Orthogonal changes of basis within either subspace preserve these norms and the constraint.

The projection decomposition gives $\|P_U Z\|_F^2 = \|Z_{\|\|}\|_F^2 + \|Z_{\|\perp}\|_F^2$ and $\|ZP_V\|_F^2 = \|Z_{\|\|}\|_F^2 + \|Z_{\perp\|}\|_F^2$. On the feasible set, $Z_{\|\|} = 0$, so the response objective equals

$$\begin{aligned}\mathcal{J}(Z) = \tfrac{1}{2}\|\tau_{\|\|}\|_F^2 + \tfrac{1}{2}\|Z_{\perp\perp} - \tau_{\perp\perp}\|_F^2 \\ + \tfrac{1}{2}\|Z_{\|\perp} - \tau_{\|\perp}\|_F^2 + \tfrac{\lambda_U}{2}\|Z_{\|\perp}\|_F^2 \\ + \tfrac{1}{2}\|Z_{\perp\|} - \tau_{\perp\|}\|_F^2 + \tfrac{\lambda_V}{2}\|Z_{\perp\|}\|_F^2.\end{aligned} \tag{48}$$

The fidelity term is coercive and strictly convex; adding nonnegative quadratic penalties preserves these properties on the closed feasible subspace. Thus a minimizer exists and is unique. This is exactly the blockwise filtering objective on the feasible subspace.

Because the four blocks are mutually orthogonal, the objective separates across them. The constraint fixes $Z_{\|\|} = 0$, while the unregularized $\perp\perp$ block satisfies

$$Z^\star_{\perp\perp} = \tau_{\perp\perp}. \tag{49}$$

For the $\|\perp$ block, first-order optimality gives

$$0 = (Z_{\|\perp} - \tau_{\|\perp}) + \lambda_U Z_{\|\perp}, \tag{50}$$

and hence

$$Z^\star_{\|\perp} = \frac{\tau_{\|\perp}}{1+\lambda_U}. \tag{51}$$

Analogously,

$$Z^\star_{\perp\|} = \frac{\tau_{\perp\|}}{1+\lambda_V}. \tag{52}$$

Therefore the unique minimizer is

$$Z^\star = \tau_{\perp\perp} + \frac{\tau_{\|\perp}}{1+\lambda_U} + \frac{\tau_{\perp\|}}{1+\lambda_V}. \tag{53}$$

Setting

$$\beta = (1+\lambda_U)^{-1}, \qquad \alpha = (1+\lambda_V)^{-1}, \tag{54}$$

recovers $\mathcal{F}_{\alpha,\beta}(\tau)$ for positive retention coefficients, while zero retention corresponds to the infinite-penalty limit.

The zero core coefficient is supplied by the preservation constraint. Without that constraint, the same separable calculation gives the core coefficient $(1+\lambda_U+\lambda_V)^{-1}$. Equal penalties imply equal cross-block coefficients; the specific default $\alpha = \beta = 1/2$ additionally sets both penalties to one in the normalized objective. These statements establish optimality for the geometric filtering objective. The experiments assess the resulting task-level performance after recovery.

A.6 Recovery Capacity, Target Conflicts, and Final Blocks

Fix a stage t and a selected layer, and write $\tau = \tau_t$, $A = A_t$, and $B = B_t$ in this analysis. Let

$$F = \mathcal{F}_{\alpha,\beta}(\tau), \qquad R = \tau BA, \qquad S_t^{\mathrm{rec}} := S_{t-1} + F + R. \tag{55}$$

Expanding the concatenated recovery objective gives

$$\mathcal{L} = \|(S_t^{\mathrm{rec}} - S_{t-1})\widehat{V}_{\le t-1}\|_F^2 + \|(S_t^{\mathrm{rec}} - \tau_t)\widehat{V}_t\|_F^2. \tag{56}$$

The first term penalizes deviations from the accumulated historical update along historical probe directions, whereas the second encourages agreement with the incoming expert along its probe directions. These objectives need not be simultaneously satisfiable.

Conflicts on shared directions. Let v be a unit vector in

$$\operatorname{span}(\widehat{V}_{\le t-1}) \cap \operatorname{span}(\widehat{V}_t).$$

Since $\widehat{V}_{\le t-1}$ and $\widehat{V}_t$ are individually orthonormal,

$$\mathcal{L} \ge \|(S_t^{\mathrm{rec}} - S_{t-1})v\|_2^2 + \|(S_t^{\mathrm{rec}} - \tau_t)v\|_2^2 \ge \frac{1}{2}\|(S_{t-1} - \tau_t)v\|_2^2. \tag{57}$$

To obtain the second inequality, let

$$z = S_t^{\mathrm{rec}} v, \qquad a = S_{t-1}v, \qquad b = \tau_t v.$$

Then

$$\|z-a\|_2^2 + \|z-b\|_2^2 = 2\left\|z - \frac{a+b}{2}\right\|_2^2 + \frac{1}{2}\|a-b\|_2^2. \tag{58}$$

Thus, if $S_{t-1}v \neq \tau_t v$ on a shared probe direction, the recovery surrogate contains unavoidable residual error even without a low-rank constraint. Equation 57 gives the contribution of this shared direction to a lower bound on the full objective.

The concatenated probe matrix has Gram matrix

$$\widehat{V}^\top \widehat{V} = \begin{bmatrix} I & \widehat{V}_{\le t-1}^\top \widehat{V}_t \\ \widehat{V}_t^\top \widehat{V}_{\le t-1} & I \end{bmatrix}. \tag{59}$$

Hence, shared directions are represented in both target terms. The number of retained columns and any normalization applied to the losses therefore affect the surrogate objective. Replacing the concatenation by an orthonormalized union without correspondingly transforming the objective would generally define a different optimization problem.

Why filtering can still matter under recovery. For the actual recovery parameterization,

$$\operatorname{rank}(R) \le \min(r, \operatorname{rank}(\tau)), \qquad \operatorname{col}(R) \subseteq \operatorname{col}(\tau). \tag{60}$$

The set of feasible corrections is therefore restricted. A finite optimization budget introduces an additional dependence on the filtered proposal from which recovery begins.

This can be contrasted with an idealized limit in which R is allowed to be an unrestricted matrix. For any fixed F, allowing R to be unrestricted makes

$$S_t^{\mathrm{rec}} = S_{t-1} + F + R$$

range over all matrices. The minimum of Eq. 56 becomes independent of F. This limit identifies the capacity constraint and finite optimization budget as the mechanisms through which filtering can influence recovery.

Recovery redistributes filtered components. For the final unit-scale increment

$$\widetilde{\tau}_t = F + R, \tag{61}$$

we have

$$P_U \widetilde{\tau}_t P_V = P_U R P_V = P_U \tau B A P_V. \tag{62}$$

The right-hand side is generally nonzero. Recovery may therefore reintroduce an $\|\|$ component and may also alter the $\|\perp$, $\perp\|$, and $\perp\perp$ components retained by the filtering proposal. Any interpretation of the final model must consequently distinguish the pre-recovery filtered update from the recovered increment.

Non-degenerate initialization of the recovery factors. Let

$$G = \frac{\partial \mathcal{L}}{\partial R}.$$

By the chain rule,

$$\nabla_A \mathcal{L} = B^\top \tau^\top G, \qquad \nabla_B \mathcal{L} = \tau^\top G A^\top. \tag{63}$$

If both factors are initialized at zero, $A = B = 0$, both gradients vanish. A non-degenerate initialization is therefore required to optimize both factors, and forms part of the recovery configuration.

A.7 Explicit Treatment of a Non-Unit Fusion Scale

The main formulation assumes unit-scale recovery and fusion; no non-unit coefficient is implicitly omitted. Suppose instead that a fixed scalar $\gamma_t^{(\ell)} = \gamma$ scales the incremental update. For a selected layer with its layer index suppressed, define

$$Z_\gamma(A, B) = S_{t-1} + \gamma\big(F + \tau BA\big), \qquad W_t^{\text{merged}} = W_0 + Z_\gamma(A, B). \tag{64}$$

If recovery is optimized for the model that will ultimately be deployed, the corresponding objective becomes

$$\begin{aligned} \mathcal{L}_\gamma &= \|Y - Z_\gamma(A, B)\widehat{V}\|_F^2 \\ &= \|\gamma(F + \tau BA)\widehat{V}_{\le t-1}\|_F^2 \\ &\quad + \|[S_{t-1} + \gamma(F + \tau BA) - \tau]\widehat{V}_t\|_F^2. \end{aligned} \tag{65}$$

An alternative implementation could optimize the unit-scale objective in Eq. 11 and apply γ only after optimization. This defines a *post-hoc scaling* procedure and is generally not equivalent to Eq. 65. Likewise, scaling the full accumulated update or assigning different scales to filtering and recovery would produce distinct algorithms. The fusion scale and its position relative to optimization jointly define the experimental configuration.

A.8 Efficient Evaluation and Offline Complexity

The filtering operator can be evaluated without explicitly materializing the dense projectors. Expanding Eq. 8 gives

$$\begin{aligned} F = \tau &- (1 - \beta)U(U^\top \tau) - (1 - \alpha)(\tau V)V^\top \\ &+ (1 - \alpha - \beta)U(U^\top \tau V)V^\top. \end{aligned} \tag{66}$$

Given the singular bases and assuming $k \le \min(d_{\text{out}}, d_{\text{in}})$, the filtering stage requires

$$O(d_{\text{out}} d_{\text{in}} k) \tag{67}$$

arithmetic operations and

$$O((d_{\text{out}} + d_{\text{in}})k) \tag{68}$$

additional storage for the bases, excluding storage of the dense update itself. The intermediate low-rank products have dimensions $k \times d_{\text{in}}$, $d_{\text{out}} \times k$, and $k \times k$. At the symmetric default $\alpha = \beta = 0.5$, the final cross term in Eq. 66 vanishes exactly.

The historical SVD used for filtering and the incoming SVD used to construct recovery probes incur separate costs. A dense full SVD has the standard arithmetic complexity

$$O(d_{\text{out}} d_{\text{in}} \min(d_{\text{out}}, d_{\text{in}})), \tag{69}$$

whereas the cost of truncated factorization depends on the numerical solver and its iteration budget. A single SVD returns both singular bases U and V. Its numerical solver and iteration budget determine the practical factorization cost.

For the unit-scale recovery objective, let

$$q = k_{\text{h}} + k_{\text{n}} \tag{70}$$

and precompute

$$E_{\text{rec}} = Y - (S_{t-1} + F)\widehat{V}. \tag{71}$$

Then each objective evaluation can be written as

$$\mathcal{L} = \|E_{\text{rec}} - (\tau B)(A\widehat{V})\|_F^2, \tag{72}$$

avoiding construction of the dense matrix τBA at every optimization step. The dominant matrix products require

$$O(d_{\text{out}} d_{\text{in}} r + d_{\text{in}} r q + d_{\text{out}} r q) \tag{73}$$

operations per iteration, with reverse-mode differentiation of comparable order. The recovery factors require $O(d_{\text{in}} r)$ storage, while the probe matrix requires $O(d_{\text{in}} q)$, in addition to the dense model matrices and intermediate products. A dense recovery correction need only be materialized once before final fusion.

This factorized evaluation describes the arithmetic cost of recovery. The efficiency study measures the full offline merging interval and separately reports inference parameters and supported-operation FLOPs.

B EXPERIMENTAL DETAILS

Experimental records. The reported experiments comprise two explicitly separated first-stage protocol cohorts. The broad benchmark and its archived supplementary analyses use $\mathcal{B}_{\mathrm{SVD}}$, which retains the established zero-history SVD convention of the benchmark implementation. The revised structural, historical-core, recovery, rank, and long-horizon controls use $\mathcal{B}_0$, under which no historical projector is constructed when the accumulated update has zero effective rank. Results from these cohorts are summarized separately and are never pooled. Each summary uses complete seeded runs as the statistical unit.

B.1 TASK SUITES AND EVALUATION

The task lists below define each suite. All experts share their backbone's pre-trained initialization. Task order is sampled once per seed and reused across all methods in the corresponding backbone–suite setting. Runs evaluate full test sets after each stage and again at the end. Batch size is 128 with eight data-loader workers.

Eight tasks (TA8). DTD, EuroSAT, GTSRB, MNIST, RESISC45, Stanford Cars, SUN397, and SVHN.

Fourteen tasks (TALL14). CIFAR-100, DTD, EuroSAT, FER2013, GTSRB, MNIST, Oxford-IIIT Pet, Oxford Flowers102, PCam, RESISC45, Stanford Cars, STL-10, SUN397, and SVHN.

Twenty tasks (TALL20). CIFAR-10, CIFAR-100, DTD, EMNIST Letters, EuroSAT, Fashion-MNIST, FER2013, Food101, GTSRB, KMNIST, MNIST, Oxford-IIIT Pet, Oxford Flowers102, PCam, Rendered SST-2, RESISC45, Stanford Cars, STL-10, SUN397, and SVHN.

Statistical unit and metrics. A complete seeded run is the unit of replication. The seed is set once and affects task order and other algorithmic randomness. Tasks, layers, and stages are aggregated within each run. In addition to main-text ACC and BWT, let

$$\mathrm{AvgSeen} = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{t}\sum_{i=1}^{t} a_{t,i}, \qquad F^{+} = \frac{1}{T-1}\sum_{i=1}^{T-1}\max\Big(0, \max_{i\le t<T} a_{t,i} - a_{T,i}\Big). \tag{74}$$

AvgSeen weights stages equally, and each stage weights its seen tasks equally. F^{+} measures nonnegative loss from the best pre-final observed accuracy on each earlier task. All reported standard deviations use `ddof`=1. Paired method differences are computed within complete seeded runs before aggregation. Differences computed from unrounded records can differ slightly from differences of rounded table means.

B.2 INITIALIZATION AND RECOVERY

First-stage initialization. We distinguish the benchmark, audited, and diagnostic first-stage settings.

Under the benchmark convention $\mathcal{B}_{\mathrm{SVD}}$, the zero historical update is processed by the inherited input-side filtering routine using the orthogonal directions returned by its SVD, and recovery is skipped at the first stage. Because the zero matrix does not identify a unique singular subspace, these directions are treated as an implementation convention rather than learned historical geometry.

Under the audited rank-zero initialization $\mathcal{B}_0$, a zero historical update is assigned zero effective historical rank and no historical projector is constructed. This convention is used for the main structural and mechanistic controls.

To isolate the first-stage effect itself, the orthogonal-bootstrap diagnostic uses matched $\mathcal{B}_0$ controls and explicitly generated orthogonal bases while keeping the subsequent BiFilt-Sym path and stage-wise random-number protocol fixed. The archived $\mathcal{B}_{\mathrm{SVD}}$ benchmark follows a different first-stage and random-number execution path from this diagnostic and is therefore not combined with it as a matched initialization comparison.

Historical and incoming probes. The inspected implementation and archived main-run configuration use filtering rank $k = 128$, historical recovery rank $k_{\mathrm{h}} = 128$, incoming probe rank $k_{\mathrm{n}} = 8$, and recovery rank $r = 64$. Both filtering and historical recovery take the leading right singular vectors of the same historical update, so $\widehat{V}_{\leq t-1} = V$. The historical-rank sweep varies both historical ranks together. Each SVD uses `torch.linalg.svd` with `full_matrices=False`; CUDA selects the `gesvd` driver. Leading singular vectors are selected after factorization. The SVD wrapper preserves the weight tensor dtype and performs no explicit precision conversion. The inspected environment uses PyTorch 2.5.1+cu124 with float32 as its default dtype; the archived Fabric configuration has no precision override.

Recovery initialization and optimization. The factor A is initialized from a zero-mean normal distribution with standard deviation 0.02, and B is initialized to zero. Thus, $\tau BA = 0$ at initialization while gradients can update B. Only the gate factors are trainable. The optimizer is Adam with learning rate 0.001, moment coefficients $(0.9, 0.999)$, numerical epsilon 10^{-8}, zero weight decay, and AMSGrad disabled. It takes 50 full-objective steps, summing unnormalized squared residuals over both probe sets and all selected layers. Each new stage creates a fresh optimizer. The recovery-disabled control sets this budget to zero. For each selected layer ℓ, unit-scale fusion gives

$$\begin{aligned}(W_t^{\mathrm{merged}})^{(\ell)} = (W_{t-1}^{\mathrm{merged}})^{(\ell)} + \mathcal{F}_{\alpha,\beta}^{(\ell)}(\tau_t^{(\ell)}) \\ + \tau_t^{(\ell)} B_t^{(\ell)} A_t^{(\ell)}.\end{aligned} \tag{75}$$

Selected layers and frozen parameters. The configuration selects leaf `nn.Linear` modules whose names contain `attn` or `fc1`. In CLIP's vision transformer these are the attention `q_proj`, `k_proj`, `v_proj`, and `out_proj` weights, plus the first MLP projection `fc1`, in each encoder block. Fusion replaces the selected weight tensors; biases, `fc2`, normalization parameters, and other unselected backbone parameters retain their pre-trained values. The text tower is unchanged. Evaluation uses batch size 128, eight workers, pinned memory, and full test sets without shuffling or dropping the final batch.

B.3 Comparison Cohorts

Benchmark coverage. All merging methods in Table 1 use seeds 42–51, giving $n = 10$ complete runs per method in each of the nine backbone–suite settings. Within a setting, methods receive the same ordered experts for a given seed and use the same test sets, evaluation checkpoints, batch configuration, and metric implementation. Consequently, method differences can be paired within seed and then summarized across the ten complete-run pairs. Bold marks the best mean among merging methods within each metric and setting, including ties. The structural controls use the smaller seed sets stated in Table 2; their summaries remain separate because they are distinct run cohorts.

Baseline implementation and selection. The baseline rows are reruns from the same benchmark codebase rather than values copied from external tables. Their operators follow the cited method definitions. Expert checkpoints, task orders, unit fusion scale, evaluation data, and aggregation are held fixed across methods. Method-specific internal hyperparameters use their cited configurations and are fixed before the main evaluation; they are not selected separately for a backbone, task suite, seed, or reported test outcome. NUFILT and BiFilt-Sym use the identical rank-64, 50-step recovery configuration described above, so their main-table comparison changes only the filter.

Controlled-comparison protocols. The revised structural controls use $\mathcal{B}_0$. The ViT-B/32 TA8 five-seed comparison uses seeds 42, 43, 44, 46, and 47; the ViT-B/16 TA8 comparison uses seeds 43, 44, and 46. The historical-core, light-recovery, and revised historical-rank controls use seeds 43, 44, and 46. The single-order ViT-B/32 TALL20 stress test uses seed 43. These controls retain the selected layers, unit fusion scale, and stated recovery configurations within their comparisons.

The archived NUFILT-versus-BiFilt recovery and structural comparisons use $\mathcal{B}_{\mathrm{SVD}}$ and seeds 43–45, with matched incoming experts and task orders. Recovery-on variants share rank 64 and 50 Adam steps; the recovery-disabled control sets the step budget to zero. The archived rank sweep varies historical filtering and recovery-probe ranks together, with other recovery settings fixed. The

coefficient follow-up uses the same local seed set. Captions identify run coverage, and standard deviations summarize complete runs rather than tasks or layers. Results from the two protocols are reported separately and are not compared through their absolute accuracy.

Distinct main and ablation cohorts. The B/32 eight-task runwise summary reports NUFILT/BiFilt-Sym BWT of $-2.7_{\pm 0.7}$/$-2.2_{\pm 0.6}$ pp. The local three-seed recovery-on cohort instead reports $-3.0_{\pm 0.8}$/$-2.4_{\pm 0.6}$ pp. These are distinct cohorts with different seed coverage. All ablation contrasts use the local paired three-seed runs.

Exploratory selection and reuse. The post-hoc coefficient grid ranks candidates by test accuracy on seed 42. The default coefficients remain fixed throughout the main evaluation. Follow-up runs use seeds 43–45, which also appear in other analyses. Twenty-three grid points were newly evaluated, while two boundary points reuse main-run records. Geometric summaries use each method's accumulated historical bases and describe the evolution of its update trajectory.

B.4 Evaluation Scope

The full benchmark covers CLIP visual classification across three backbones and three task suites. Targeted controls evaluate four filter structures on two TA8 backbones, plus one B/32 TALL20 order. Supplementary analyses examine paired comparisons, trajectories, coefficient sensitivity, block geometry, taskwise performance, and merging resources. Geometric measurements follow method-specific histories. Resource measurements cover offline merging time and memory, together with parameter counts and supported-operation FLOPs. The theoretical analysis characterizes layerwise parameter-update geometry under fixed projectors and explicit input-alignment conditions.

C Additional Structural Controls

The broad benchmark retains the established $\mathcal{B}_{\mathrm{SVD}}$ convention, whereas the structural controls in this section use the audited $\mathcal{B}_0$ initialization. The two archived cohorts follow different execution paths and are therefore reported separately; their absolute accuracy values are not used as an initialization comparison. First-stage effects are instead evaluated by the matched orthogonal-bootstrap diagnostic in Section C.1, which keeps the subsequent code path and random-number protocol fixed.

Unless otherwise stated, the remaining TA8 structural controls use ViT-B/32 and seeds 43, 44, and 46.

C.1 First-Stage Orthogonal Bootstrap Diagnostic

The broad benchmark retains $\mathcal{B}_{\mathrm{SVD}}$ for consistency with the established benchmark implementation, whereas the structural controls use the audited rank-zero initialization $\mathcal{B}_0$. Because these cohorts follow different archived execution paths, their absolute accuracies do not provide a controlled initialization comparison. We therefore run a separate matched diagnostic that changes only the first-stage treatment while keeping the subsequent BiFilt-Sym path and stage-wise random-number protocol fixed.

We use ViT-B/32 TA8 with task-order seeds 42, 43, 44, 46, and 47. For each order, the rank-zero control directly incorporates the first expert. The orthogonal-bootstrap variants instead apply the first-stage input-side filter using three independently generated orthogonal bases, denoted b_0, b_1, and b_2. All subsequent stages use the same BiFilt-Sym configuration: 50-step rank-64 recovery, historical rank 128, and incoming rank 8.

The ACC improvement is consistent across both sources of variation: all five task-order seeds improve after averaging over bases, and all 15 individual basis–order runs have positive paired ACC differences. The three basis choices yield similar five-seed gains ($+0.3$, $+0.3$, and $+0.4$ pp), with only 0.1 pp separating their mean ACC values. In contrast, BWT changes are small and mixed in sign; the mean difference rounds to 0.0 pp. These results indicate that, in this setting, the benefit of a nondegenerate first-stage projection does not depend on one particular arbitrary orthogonal basis and is consistent with a bootstrap-regularization effect.

Table 4: **First-stage orthogonal-bootstrap diagnostic on ViT-B/32 TA8.** Each basis row averages the same five task-order seeds (42, 43, 44, 46, 47). Differences are paired against the corresponding rank-zero control. "Orth mean" first averages the three bases within each task order and then averages across task orders. Differences are computed before rounding; all displayed results have one decimal.

First-stage rule	ACC (%) ↑	BWT (pp) ↑	ΔACC	ΔBWT
Rank-0	83.8	-1.9	0.0	0.0
Orth b_0	84.1	-1.9	+0.3	0.0
Orth b_1	84.1	-1.9	+0.3	-0.1
Orth b_2	**84.2**	-1.9	**+0.4**	0.0
Orth mean	84.1	-1.9	+0.3	0.0

This diagnostic is intentionally limited in scope. It evaluates ViT-B/32 TA8 and five task orders and therefore does not establish that orthogonal bootstrap improves every backbone or task suite. Moreover, the archived $\mathcal{B}_{\mathrm{SVD}}$ benchmark follows a different first-stage and random-number execution path from this matched diagnostic. We consequently use the diagnostic to characterize the effect of structured first-stage projection, not to attribute the absolute benchmark difference between $\mathcal{B}_{\mathrm{SVD}}$ and $\mathcal{B}_0$ to initialization alone.

The remaining structural controls therefore use the audited $\mathcal{B}_0$ initialization. Unless otherwise stated, TA8 summaries use ViT-B/32 and seeds 43, 44, and 46.

Table 5: B/32 TALL20 structural stress test under the audited $\mathcal{B}_0$ initialization, seed 43 and one task order. Each entry is an individual-run result; OldAcc-final averages final accuracy on tasks preceding the last stage.

Filter	ACC	BWT	AvgSeen	Acquire	OldAcc-final
RIGHT	69.6	-10.7	77.4	79.8	70.4
LEFT	71.3	-9.0	78.6	79.8	72.1
HARD	**72.4**	**-8.1**	78.7	80.1	**73.2**
BiFilt-Sym	71.8	-9.2	**78.8**	**80.6**	72.6

Table 6: Historical-core ($\|\|\|$) retention under the audited $\mathcal{B}_0$ initialization on B/32 TA8; mean $\pm$ sample SD over three complete runs. Other block weights follow BiFilt-Sym.

Recovery	Core weight	ACC (%)	BWT (pp)
Off	0	$80.8_{\pm 1.1}$	$-2.1_{\pm 0.6}$
	0.25	$79.6_{\pm 1.2}$	$-3.3_{\pm 0.7}$
	$1/3$	$78.9_{\pm 1.2}$	$-3.9_{\pm 0.7}$
50 steps, rank 64	0	$83.8_{\pm 0.3}$	$-2.0_{\pm 0.3}$
	0.25	$83.6_{\pm 0.4}$	$-2.1_{\pm 0.2}$
	$1/3$	$83.6_{\pm 0.4}$	$-2.1_{\pm 0.2}$

D Additional Benchmark Analyses

All results in this section belong to the archived $\mathcal{B}_{\mathrm{SVD}}$ benchmark cohort and retain its zero-history SVD convention. They are reported separately from the audited $\mathcal{B}_0$ structural controls in Section 5.3 and Appendix C; absolute ACC or BWT differences between these cohorts are not interpreted as initialization effects. Seed coverage and comparison arms are stated for each analysis. All $\pm$ values are sample standard deviations across complete runs (ddof$= 1$). Bold marks the best comparable mean (ties included): higher accuracy/BWT and lower forgetting/cost. Paired tables compare methods within rows. Trajectories, contrasts, and geometry are descriptive summaries. A dash marks an unavailable measurement or an explicitly undefined quantity.

Table 7: B/32 TA8 recovery and historical-rank sensitivity under the audited $\mathcal{B}_0$ initialization over seeds 43, 44, and 46. Light recovery uses 10 steps and rank 16; the rank sweep fixes BiFilt-Sym with 50-step, rank-64 recovery.

Control	Setting	ACC (%)	BWT (pp)
Light recovery	RIGHT	$80.0_{\pm 1.1}$	$-2.6_{\pm 0.9}$
	LEFT	$81.8_{\pm 0.4}$	$-2.9_{\pm 0.6}$
	BiFilt-Sym	$\mathbf{81.9}_{\pm 0.8}$	$\mathbf{-1.8}_{\pm 0.5}$
Historical rank	64	**84.0**	-2.5
	128	83.8	-2.0
	256	82.5	**-1.6**

D.1 Complete Main-Run Metrics

Table 8: BiFilt-Sym main results, each over seeds 42–51 ($n = 10$). ACC and AvgSeen are in percent; BWT and F^+ are in pp. Bold compares backbones within each task suite.

Setting	ACC	BWT	AvgSeen	F^+
B32_TA8	$84.0_{\pm 0.2}$	$-2.2_{\pm 0.6}$	$86.5_{\pm 3.4}$	$2.2_{\pm 0.6}$
B32_TALL14	$79.2_{\pm 0.2}$	$-4.9_{\pm 0.7}$	$83.2_{\pm 2.4}$	$5.0_{\pm 0.7}$
B32_TALL20	$72.6_{\pm 0.8}$	$-8.3_{\pm 2.0}$	$79.5_{\pm 1.2}$	$8.4_{\pm 2.0}$
B16_TA8	$87.5_{\pm 0.1}$	$-1.3_{\pm 0.4}$	$88.9_{\pm 3.1}$	$1.3_{\pm 0.4}$
B16_TALL14	$84.0_{\pm 0.3}$	$-3.0_{\pm 0.6}$	$86.4_{\pm 2.3}$	$3.0_{\pm 0.6}$
B16_TALL20	$79.6_{\pm 0.8}$	$-6.2_{\pm 1.6}$	$84.4_{\pm 1.1}$	$6.3_{\pm 1.6}$
L14_TA8	$\mathbf{91.8}_{\pm 0.1}$	$\mathbf{-0.9}_{\pm 0.3}$	$\mathbf{92.7}_{\pm 2.1}$	$\mathbf{0.9}_{\pm 0.3}$
L14_TALL14	$\mathbf{89.6}_{\pm 0.1}$	$\mathbf{-1.8}_{\pm 0.3}$	$\mathbf{91.1}_{\pm 1.7}$	$\mathbf{1.8}_{\pm 0.3}$
L14_TALL20	$\mathbf{85.5}_{\pm 0.5}$	$\mathbf{-4.4}_{\pm 0.6}$	$\mathbf{89.1}_{\pm 1.1}$	$\mathbf{4.5}_{\pm 0.6}$

D.2 Paired Main Comparisons

All nine NUFILT–BiFilt-Sym comparisons use the same ten seeds and task orders within each setting. Table 10 reports the mean within-seed differences at the one-decimal precision of the main benchmark table.

The detailed runwise analysis below focuses on ViT-B/32 with eight tasks. Each seed fixes the task order and other algorithmic randomness, so the complete seeded run is the pairing unit. We report paired mean differences and the number of runs in which BiFilt-Sym improves the corresponding metric.

Table 9: Paired main-table differences, BiFilt-Sym minus NUFILT. Both methods use the same seeds 42–51 and task orders in every setting ($n = 10$ pairs). Values are differences of the one-decimal means reported in Table 1; positive values favor BiFilt-Sym.

Backbone	Tasks	ΔACC (pp)	ΔBWT (pp)
ViT-B/32	8	+0.4	+0.5
ViT-B/32	14	+1.2	+0.8
ViT-B/32	20	+1.6	+0.6
ViT-B/16	8	+0.2	+0.3
ViT-B/16	14	+0.9	+0.5
ViT-B/16	20	+1.5	+0.9
ViT-L/14	8	+0.2	+0.2
ViT-L/14	14	+0.4	+0.2
ViT-L/14	20	+0.8	+0.2

Table 10: Paired run analysis on ViT-B/32 with eight tasks, seeds 42–51. Entries are mean ± sample SD. Differences are BiFilt-Sym minus NUFILT; lower F^+ is better. Wins count paired runs in which BiFilt-Sym improves the corresponding metric.

Metric	NUFILT	BiFilt-Sym	Difference	Wins
ACC	$83.6_{\pm 0.2}$	$\mathbf{84.0}_{\pm 0.2}$	$+0.4$	$10/10$
BWT	$-2.7_{\pm 0.7}$	$\mathbf{-2.2}_{\pm 0.6}$	$+0.5$	$10/10$
AvgSeen	$86.3_{\pm 3.4}$	$\mathbf{86.5}_{\pm 3.4}$	$+0.1$	$9/10$
F^+	$2.7_{\pm 0.7}$	$\mathbf{2.2}_{\pm 0.6}$	-0.5	$10/10$

Table 10 summarizes the paired runwise results. BiFilt-Sym improves ACC by 0.4 pp and BWT by 0.5 pp on average, with positive paired differences in all ten runs. AvgSeen increases by 0.1 pp in nine of ten runs, while nonnegative forgetting F^+ decreases by 0.5 pp in all ten runs. Figure 4 shows the corresponding runwise and taskwise outcomes.

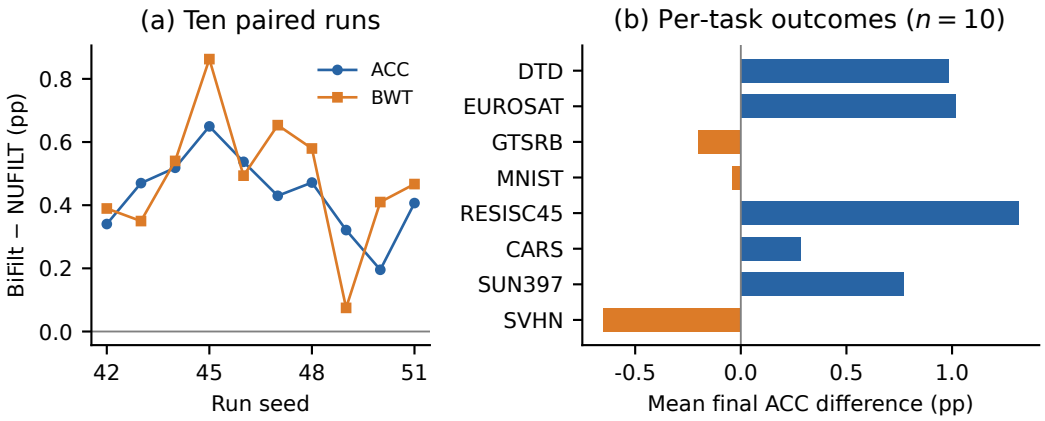

Figure 4: B/32 eight-task main cohort. (a) ACC and BWT differences are positive in all ten paired runs. (b) Final per-task mean gains are heterogeneous; SVHN loses accuracy in every seed. Panels reuse the same runs.

Table 11: Paired B/32 eight-task records from the ten-seed main cohort. All rows use the same main-run protocol.

Seed	ACC_N	ACC_B	ΔACC	BWT_N	BWT_B	ΔBWT
42	83.7	**84.0**	0.3	-1.9	**-1.5**	0.4
43	83.3	**83.8**	0.5	-2.1	**-1.7**	0.4
44	84.0	**84.5**	0.5	-3.0	**-2.4**	0.5
45	83.3	**84.0**	0.7	-3.8	**-2.9**	0.9
46	83.4	**84.0**	0.5	-2.6	**-2.1**	0.5
47	83.6	**84.0**	0.4	-3.1	**-2.4**	0.7
48	83.5	**84.0**	0.5	-2.4	**-1.8**	0.6
49	83.6	**83.9**	0.3	-2.2	**-2.1**	0.1
50	83.9	**84.1**	0.2	-2.1	**-1.7**	0.4
51	83.6	**84.1**	0.4	-3.7	**-3.2**	0.5

D.3 Archived Recovery and Structural Controls

The archived $\mathcal{B}_{\mathrm{SVD}}$ ablation cohort consists of seeds 43–45. Recovery-on baselines use the local resource-analysis reruns. The recovery-disabled comparison sets the optimization budget to zero in both methods. Let g_0 and g_1 denote the paired BiFilt-Sym-minus-NUFILT difference with recovery

Table 12: Longer B/32 suites: seed 42 only. Entries are individual-run measurements.

Setting	ACC_N	ACC_B	ΔACC	BWT_N	BWT_B	ΔBWT
B32_TALL14	77.7	**79.1**	1.4	-6.7	**-5.7**	1.0
B32_TALL20	72.2	**73.5**	1.3	-6.0	**-5.6**	0.4

off and on. The interaction $g_1 - g_0$ is computed within seed before averaging. The paired contrasts are reported in Table 24. Changes in BWT also reflect changes in task-arrival reference accuracies.

Table 13: Recovery factorial on three paired seeds. Off denotes zero recovery steps. Bold marks the best mean within each recovery condition; lower F^+ is better.

Filter	Recovery	ACC	BWT	AvgSeen	F^+
NUFILT	Off	$80.4_{\pm 1.2}$	$\mathbf{-1.9}_{\pm 0.8}$	$84.0_{\pm 1.9}$	$\mathbf{2.1}_{\pm 0.7}$
NUFILT	On	$83.5_{\pm 0.4}$	$-3.0_{\pm 0.8}$	$86.6_{\pm 2.5}$	$3.0_{\pm 0.8}$
BiFilt-Sym	Off	$\mathbf{81.7}_{\pm 0.7}$	$-2.3_{\pm 0.8}$	$\mathbf{85.2}_{\pm 2.3}$	$2.4_{\pm 0.8}$
BiFilt-Sym	On	$\mathbf{84.1}_{\pm 0.4}$	$\mathbf{-2.4}_{\pm 0.6}$	$\mathbf{86.7}_{\pm 2.6}$	$\mathbf{2.4}_{\pm 0.6}$

Table 14: Extended structural controls, seeds 43–45. All methods use recovery. $\perp\|$-relaxed reuses the coefficient follow-up runs. Bold marks the best unrounded mean in each metric column; values are displayed to one decimal.

Filter	(α, β)	ACC	BWT
Right-only	$0, 1$	$83.5_{\pm 0.4}$	$-3.0_{\pm 0.8}$
Left-only	$1, 0$	$\mathbf{84.1}_{\pm 0.5}$	$-2.1_{\pm 0.6}$
Bilateral-hard	$0, 0$	$83.7_{\pm 0.5}$	$\mathbf{-2.0}_{\pm 0.6}$
$\perp\|$-relaxed	$0.5, 1$	$84.0_{\pm 0.4}$	$-2.8_{\pm 0.8}$
BiFilt-Sym	$0.5, 0.5$	$84.1_{\pm 0.4}$	$-2.4_{\pm 0.6}$

The shared-soft product is $(P_U^\perp + cP_U)\tau(P_V^\perp + cP_V)$ at $c = 0.5$, retaining $0.25\tau_{\|\|}$ in addition to the default cross-block weights. It is distinct from BiFilt-Sym. Left-only ties BiFilt-Sym in ACC at one-decimal precision and is 0.3 pp higher in BWT on average; it has higher BWT in all three seeds and higher ACC in two. BiFilt-Sym exceeds bilateral-hard in ACC by 0.4 pp but has lower BWT by 0.3 pp. Thus, retaining cross-block information increases ACC relative to hard bilateral removal, while the one-sided and bilateral choices produce different ACC–BWT trade-offs.

D.4 Continual Trajectories and Per-Task Outcomes

The stagewise differences below average the three-seed archived $\mathcal{B}_{\mathrm{SVD}}$ ablation cohort. Within that cohort the first-stage routine is shared, and old-task accuracy starts at stage 2. Later stages average over the task identities presented at that position in each seeded sequence.

D.4.1 Final per-task performance

The next table instead uses the ten-seed main cohort. Task results decompose the final macro-average into its eight dataset components. Five of eight task means improve; SVHN worsens in every paired seed.

D.5 Exploratory Coefficient Sensitivity

The 25-point grid uses seed 42. Twenty-three points were newly evaluated; the right-only and default points reuse main-run records. The highest-ACC grid point is $(0.75, 1)$; 13 of 25 points are within 0.1 pp of its ACC. Follow-up candidates are that pair and the two nearest Manhattan-grid neighbors,

Table 15: BiFilt-Sym-minus-NUFILT trajectory differences (pp), seeds 43–45. Entries are rounded mean differences across complete runs.

Stage	Seen-task ACC	Old-task ACC	New-task ACC
1	0.0	-	0.0
2	-0.1	+0.1	-0.3
3	0.0	0.0	0.0
4	+0.1	+0.1	+0.1
5	+0.1	+0.1	0.0
6	+0.2	+0.2	0.0
7	+0.4	+0.4	+0.3
8	+0.5	+0.6	+0.2

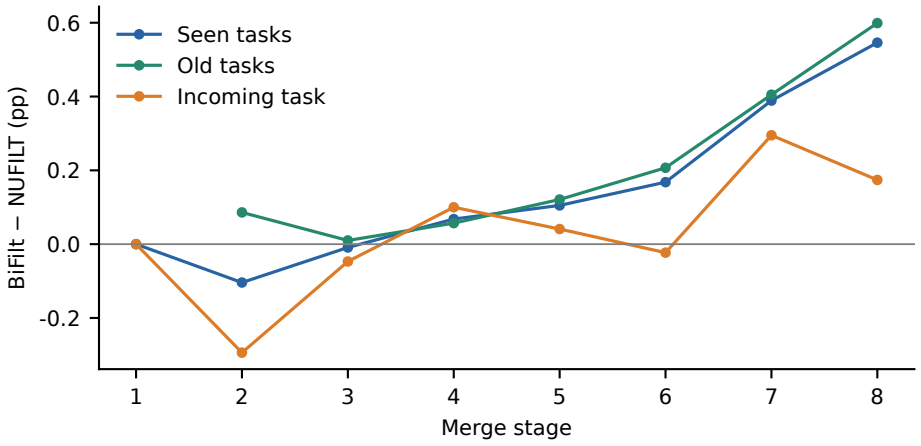

Figure 5: Stagewise paired differences on three seeds. Curves show mean differences across complete runs.

Table 16: Final per-task results over ten paired seeds. SD is computed across the ten paired differences.

Task	NUFILT	BiFilt-Sym	ΔACC	SD(Δ)	Wins
dtd	64.8	**65.8**	1.0	0.8	10/10
eurosat	93.1	**94.1**	1.0	0.3	10/10
gtsrb	**92.5**	92.3	-0.2	0.4	5/10
mnist	**98.9**	**98.9**	0.0	0.1	2/10
resisc45	84.9	**86.2**	1.3	0.4	10/10
stanford-cars	72.0	**72.2**	0.3	0.5	7/10
sun397	69.2	**70.0**	0.8	0.2	10/10
svhn	**93.2**	92.5	-0.7	0.2	0/10

with ties ordered by β then α. Follow-up runs reuse seeds 43–45 from the experimental cohort. The grid is an exploratory sensitivity analysis; the symmetric coefficients stay fixed in the main runs.

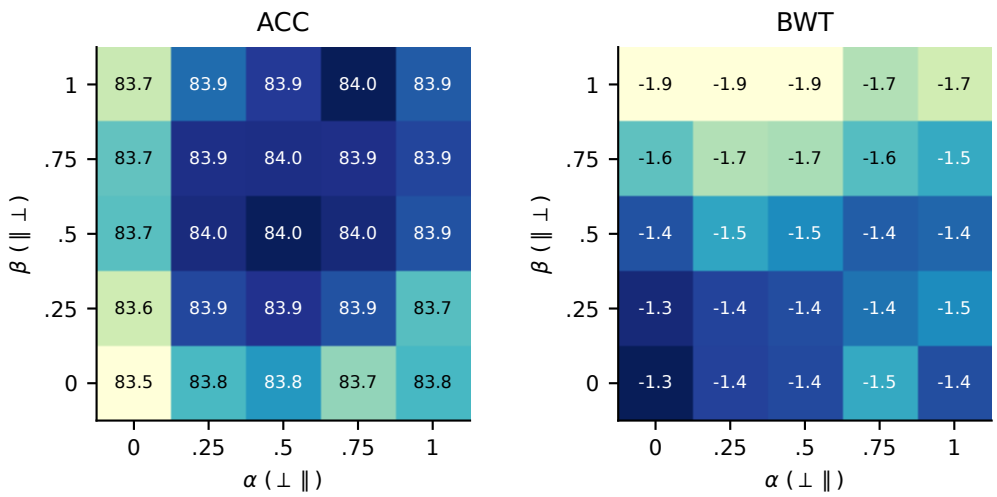

Figure 6: Exploratory coefficient grid. The default and NUFILT boundary reuse main-run records; the other cells are newly evaluated points.

Table 17: Seed-42 ACC grid. Rows control $\|\perp$ and columns control $\perp\|$. Bold marks the highest unrounded value; cells are displayed to one decimal.

$\beta\backslash\alpha$	0	0.25	0.5	0.75	1
0	83.5	83.8	83.8	83.7	83.8
0.25	83.6	83.9	83.9	83.9	83.7
0.5	83.7	84.0	84.0	84.0	83.9
0.75	83.7	84.0	84.0	83.9	83.9
1	83.7	83.9	83.9	**84.0**	83.9

Table 18: Seed-42 BWT grid. Rows control $\|\perp$ and columns control $\perp\|$. Bold marks the highest unrounded value; cells are displayed to one decimal.

$\beta\backslash\alpha$	0	0.25	0.5	0.75	1
0	**-1.3**	-1.4	-1.4	-1.5	-1.4
0.25	-1.3	-1.4	-1.4	-1.4	-1.5
0.5	-1.4	-1.5	-1.5	-1.4	-1.4
0.75	-1.6	-1.7	-1.7	-1.6	-1.5
1	-1.9	-1.9	-1.9	-1.7	-1.7

The default has the highest mean ACC and BWT among these four follow-up configurations. Both the grid ranking and follow-up are post-hoc exploratory analyses based on test accuracy. They characterize local sensitivity within the tested configurations and do not alter the fixed main-evaluation coefficients.

D.6 Block Geometry Before and After Recovery

Nine checkpoint sequences cover NUFILT, BiFilt-Sym, and bilateral-hard on three seeds. For each method's own trajectory, the analysis covers stages 2–8 and all target linear layers. For block b, the per-run summaries are

$$e_b = \frac{\sum_{t,\ell} \|\tau_{t,b}^{(\ell)}\|_F^2}{\sum_{t,\ell} \|\tau_t^{(\ell)}\|_F^2}, \qquad \rho_b^{\text{post}} = \sqrt{\frac{\sum_{t,\ell} \|[(W_t^{\text{merged}})^{(\ell)} - (W_{t-1}^{\text{merged}})^{(\ell)}]_b\|_F^2}{\sum_{t,\ell} \|\tau_{t,b}^{(\ell)}\|_F^2}}. \tag{76}$$

Table 19: Coefficient sensitivity on seeds 43–45. Candidates come from the seed-42 test-accuracy grid; the default reuses local recovery-on runs. Bold marks the best mean in each metric column.

(α, β)	ACC (%)	BWT (pp)	ΔACC vs. default
$(0.5, 0.5)$	$\mathbf{84.1}_{\pm 0.4}$	$\mathbf{-2.4}_{\pm 0.6}$	**0.0**
$(0.75, 1)$	$84.0_{\pm 0.5}$	$-2.8_{\pm 0.9}$	-0.1
$(0.75, 0.75)$	$84.0_{\pm 0.4}$	$-2.6_{\pm 0.8}$	-0.1
$(0.5, 1)$	$84.0_{\pm 0.4}$	$-2.8_{\pm 0.8}$	-0.1

The pre-recovery ratio replaces the checkpoint increment in the numerator by the filtered proposal. Each table averages these ratios across seeds after pooling within a seed. The ratios compare Frobenius norms and can exceed one. Each method uses the historical bases of its own accumulated model, so the summaries describe its realized trajectory rather than a shared fixed basis.

Table 20: Pooled-within-run block summaries, averaged across three seeds per method. Pre-recovery numerical zeros are rounded; their small residuals reflect numerical precision.

Quantity	Method	$\|\|\|$	$\|\|\perp$	$\perp\|\|$	$\perp\perp$
Energy (%)	NUFILT	10.3	16.7	24.5	48.5
Energy (%)	BiFilt-Sym	10.4	15.6	25.5	48.5
Energy (%)	HARD	9.8	15.2	25.6	49.4
ρ^{pre}	NUFILT	0.0	1.0	0.0	1.0
ρ^{pre}	BiFilt-Sym	0.0	0.5	0.5	1.0
ρ^{pre}	HARD	0.0	0.0	0.0	1.0
ρ^{post}	NUFILT	0.0	1.4	0.0	1.3
ρ^{post}	BiFilt-Sym	0.2	1.3	0.3	1.4
ρ^{post}	HARD	0.1	1.1	0.0	1.4

BiFilt-Sym's mean input-update energy is approximately 10.4%, 15.6%, 25.5%, and 48.5% in $\|\|\|$, $\|\perp$, $\perp\|$, and $\perp\perp$ (rounded values need not sum exactly to 100%). Its post-recovery $\perp\|$ ratio is higher than those of NUFILT and hard filtering. However, its $\|\|\|$ ratio is 0.2 despite a zero filtering weight, and bilateral-hard has $\|\perp$ ratio 1.1 despite removing that block before recovery. These measurements show how recovery redistributes components beyond the original filtering weights along each method's trajectory.

D.7 Historical-Rank Sensitivity

The archived $\mathcal{B}_{\mathrm{SVD}}$ rank sweep uses three paired seeds per rank; rank 128 reuses the local recovery-on runs. Recovery rank and incoming probe rank remain 64 and 8. The historical-rank setting controls both the first-stage SVD and the later historical subspaces.

Table 21: Historical-rank sensitivity under the archived $\mathcal{B}_{\mathrm{SVD}}$ protocol. All four ranks yield positive mean ACC and BWT differences relative to NUFILT within this cohort.

Rank	ACC_N	ACC_B	ΔACC	BWT_N	BWT_B
32	$82.6_{\pm 0.2}$	$\mathbf{83.1}_{\pm 0.4}$	+0.5	$-4.8_{\pm 1.2}$	$\mathbf{-4.3}_{\pm 1.3}$
64	$83.4_{\pm 0.4}$	$\mathbf{83.9}_{\pm 0.4}$	+0.5	$-3.7_{\pm 1.2}$	$\mathbf{-3.2}_{\pm 0.9}$
128	$83.5_{\pm 0.4}$	$\mathbf{84.1}_{\pm 0.4}$	+0.5	$-3.0_{\pm 0.8}$	$\mathbf{-2.4}_{\pm 0.6}$
256	$82.7_{\pm 0.3}$	$\mathbf{83.0}_{\pm 0.1}$	+0.4	$-3.0_{\pm 0.8}$	$\mathbf{-2.3}_{\pm 0.8}$

D.8 Merging Resources and Inference Structure

Methods were alternated on the same compute host for seeds 43–45. Timing uses synchronized merging intervals, including SVD, filtering, recovery, and folding, while excluding model loading,

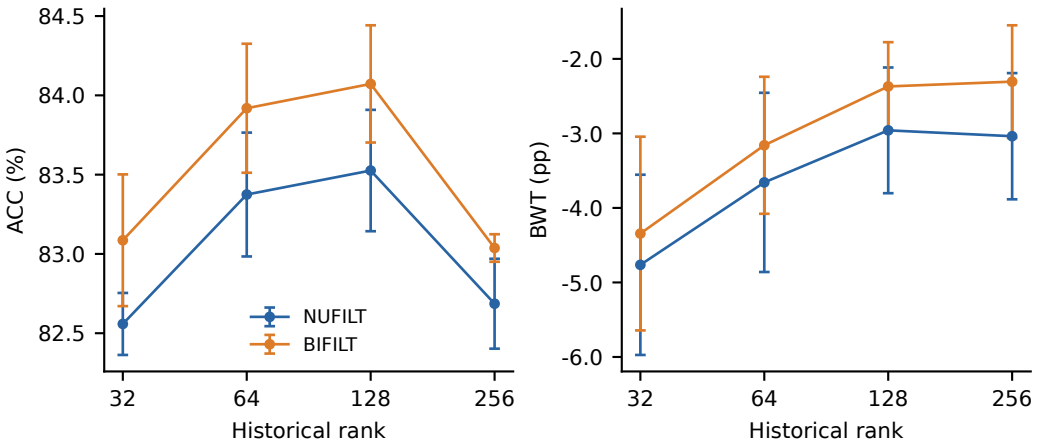

Figure 7: Historical-rank sweep under the archived $\mathcal{B}_{\mathrm{SVD}}$ protocol on three seeds. Error bars are sample SD. BiFilt-Sym has higher mean ACC and BWT at all four ranks, but the two metrics favor different ranks. Rank 128 reuses the local recovery-on runs.

saving, and evaluation. Other jobs shared host CPU and I/O resources. Timing repetitions use the same three seeded runs per method.

Table 22: Eight-stage merge cost. Time is mean ± sample SD across three runs; memory is the mean of each run's peak allocated GPU memory.

Method	Time (s)	Peak GiB	Extra params	Extra counted FLOPs
NUFILT	$218.6_{\pm 1.1}$	**2.7**	**0**	**0**
BiFilt-Sym	$\mathbf{214.9}_{\pm 0.8}$	2.9	**0**	**0**

The vision encoder contains 87,456,000 parameters. For input shape $1 \times 3 \times 224 \times 224$, the profiler counts 8,729,323,008 supported-operation FLOPs both before and after merging. The count covers operations supported by the profiler. Checkpoint keys and tensor shapes match the original CLIP backbone; no inference module is added and the text tower is unchanged. BiFilt-Sym uses about 0.2 GiB more peak memory. The observed mean merge times differ by 1.7%, indicating comparable cost in this shared-host measurement. End-to-end deployment latency was not measured.

D.9 Archived Recovery and Block-Geometry Analyses

The analyses in this subsection belong to the archived $\mathcal{B}_{\mathrm{SVD}}$ cohort and use seeds 43–45. They are distinct from the audited $\mathcal{B}_0$ structural controls in the main text and Appendix C. Results from the two cohorts are not pooled, and differences in their absolute ACC or BWT values are not interpreted as initialization effects.

Filtering before recovery. With recovery disabled, BiFilt-Sym improves ACC over NUFILT by 1.3 pp, while BWT changes by -0.4 pp. This comparison isolates the filtering proposal. With recovery enabled, the differences are +0.5 pp ACC and +0.6 pp BWT (Table 23). Recovery raises ACC by 3.1 pp for NUFILT and 2.4 pp for BiFilt-Sym, yielding an ACC interaction of -0.8 pp (Table 24). Recovery therefore narrows the ACC gap and reverses the BWT comparison.

Cross-block retention under shared recovery. With recovery fixed, retaining half of each cross block raises ACC by 0.4 pp over bilateral-hard filtering and lowers BWT by 0.3 pp. The left-only rule ties BiFilt-Sym in ACC at one-decimal precision and has 0.3 pp higher BWT (Table 14). Partial retention can therefore improve final accuracy, although symmetric weighting is not uniformly best. The main benchmark uses the same symmetric configuration throughout.

Table 23: Recovery and cross-block retention under the archived $\mathcal{B}_{\mathrm{SVD}}$ protocol on B/32 eight tasks, seeds 43–45. **Bold** marks the best mean among the shown methods within each recovery condition. Complete boundary comparisons appear in Table 14.

Filter	Recovery	(α, β)	ACC (%)	BWT (pp)
NUFILT	Off	$0, 1$	$80.4_{\pm 1.2}$	$\mathbf{-1.9}_{\pm 0.8}$
BiFilt-Sym	Off	$0.5, 0.5$	$\mathbf{81.7}_{\pm 0.7}$	$-2.3_{\pm 0.8}$
NUFILT	On	$0, 1$	$83.5_{\pm 0.4}$	$-3.0_{\pm 0.8}$
BiFilt-Sym	On	$0.5, 0.5$	$\mathbf{84.1}_{\pm 0.4}$	$\mathbf{-2.4}_{\pm 0.6}$

Table 24: Paired recovery contrasts under the archived $\mathcal{B}_{\mathrm{SVD}}$ protocol on seeds 43–45 (mean $\pm$ sample SD, pp). Filter effects are BiFilt-Sym minus NUFILT; recovery effects are on minus off. The interaction is the change in the filter effect after recovery. Contrasts are computed within seed before averaging.

Contrast	ACC	BWT
Filter effect, recovery off	$+1.3_{\pm 0.6}$	$-0.4_{\pm 0.1}$
Filter effect, recovery on	$+0.5_{\pm 0.1}$	$+0.6_{\pm 0.3}$
Recovery interaction	$-0.8_{\pm 0.7}$	$+1.0_{\pm 0.3}$
Recovery effect, NUFILT	$+3.1_{\pm 0.9}$	$-1.0_{\pm 0.3}$
Recovery effect, BiFilt-Sym	$+2.4_{\pm 0.3}$	$0.0_{\pm 0.3}$

Block geometry after recovery. Recovery changes the proposal's block structure (Figure 8). For BiFilt-Sym, the pre-recovery weights $(0, 0.5, 0.5, 1)$ for $\|\|\|/\|\|\perp/\perp\|/\perp\perp$ become final norm ratios of 0.2/1.3/0.3/1.4. The nonzero $\|\|\|$ ratio confirms that the final increment leaves the proposal's core-preserving set. Bilateral-hard filtering likewise yields a $\|\perp$ ratio of 1.1 after removing that block in the proposal. Ratios pool squared norms over layers and stages within each run and then average over seeds using each method's historical bases. Incoming-update energy for BiFilt-Sym is 10.4%, 15.6%, 25.5%, and 48.5% across the four blocks. The cross blocks therefore contain 41.1% of the incoming energy, while the controlled filtering comparisons isolate their behavioral effect.

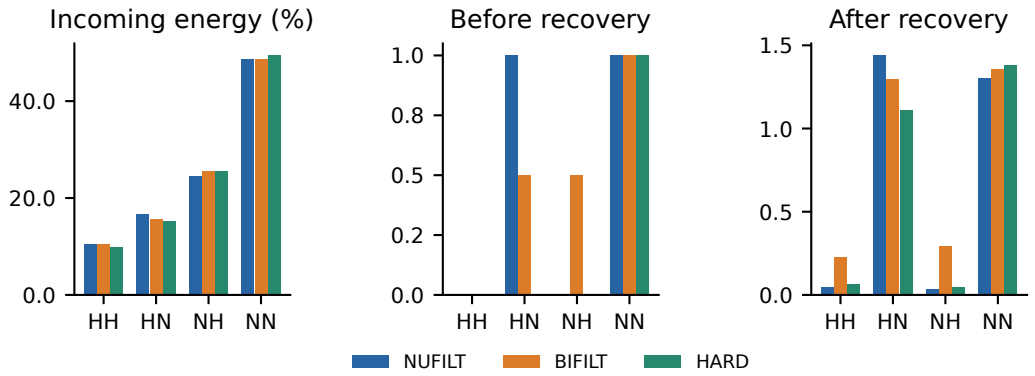

Figure 8: **Block energy and recovery effects under archived $\mathcal{B}_{\mathrm{SVD}}$.** Cross blocks contain 41.1% of incoming energy. Pre-recovery ratios verify filter coefficients; the nonzero final ||| ratio shows departure from core preservation, while ratios above one indicate amplification relative to the incoming update.