THE STABILITY TRAP IN MULTIMODAL FOUNDATION MODEL CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual adaptation keeps multimodal foundation models useful as data evolve, yet historical protection can impose an avoidable cost on new-task learning—a *stability trap*. Focusing on vision–language models, we identify and reduce protection-induced plasticity loss: adaptation recoverable by relaxing historical constraints while keeping losses in historical and pretrained performance within specified tolerances. We measure this loss through paired learning runs that start from identical model and optimizer states, vary only the retained protection, and jointly assess new-task learning and knowledge retention. We introduce Budgeted Functional Protection (BFP), which periodically reallocates a fixed number of protected parameter directions according to their estimated influence on model outputs, leaving the remaining directions available for learning. On VLCL under a common training budget, BFP improves image-to-text and text-to-image Recall@1 over Pi-CCA by 1.11 and 1.06 percentage points, respectively, while reducing pretrained accuracy drop from 3.02 to 1.37 points. The results support selective protection as a way to improve learning and retention, with the paired diagnostic measuring the adaptation cost of the retained constraints.

1 INTRODUCTION

Pretrained vision–language models (VLMs) support visual recognition and image–text retrieval through transferable cross-modal representations (Radford et al., 2021). As data evolve, these multimodal foundation models must acquire new knowledge without retraining on the full history, while retaining learned tasks and pretrained zero-shot capabilities.

Figure 1 illustrates the tension: protecting existing knowledge can suppress changes needed for new learning. ZSCL preserves pretrained behavior through distillation and weight ensembling (Zheng et al., 2023); Pi-CCA maintains cross-modal alignment statistics without replay (Zhang et al., 2026). C-CLIP improves new-domain retrieval through contrastive knowledge consolidation, highlighting that reduced forgetting can accompany inadequate learning (Liu et al., 2025). These advances motivate joint evaluation of retention and adaptation, but leave a question: how much adaptation loss is necessary to preserve knowledge, and how much comes from the protection mechanism?

This distinction matters when constraints restrict update directions. Orthogonal Gradient Descent protects historical outputs (Farajtabar et al., 2020); KeepLoRA extends subspace protection to general and task-specific VLM knowledge (Luo et al., 2026). Directions important to past knowledge may also support future learning. FS-DGPM relaxes protection adaptively (Deng et al., 2021), while space decoupling treats stability- and plasticity-related directions differently (Zhao et al., 2023). We ask how much adaptation is lost *because of protection* and can be recovered without materially compromising historical or pretrained capabilities. Since plasticity also deteriorates during ordinary continual optimization (Dohare et al., 2024), attribution requires a controlled constraint change and joint adaptation–retention measurement.

We call this avoidable suppression a *stability trap* and quantify it through *protection-induced plasticity loss* (PIPL): adaptation recoverable by relaxing historical protection at a fixed learned state and training budget, subject to historical and pretrained retention tolerances. Paired learning runs start from the same model and optimizer checkpoint with retained or relaxed protection. We introduce *Budgeted Functional Protection* (BFP), which reallocates a fixed number of protected directions

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

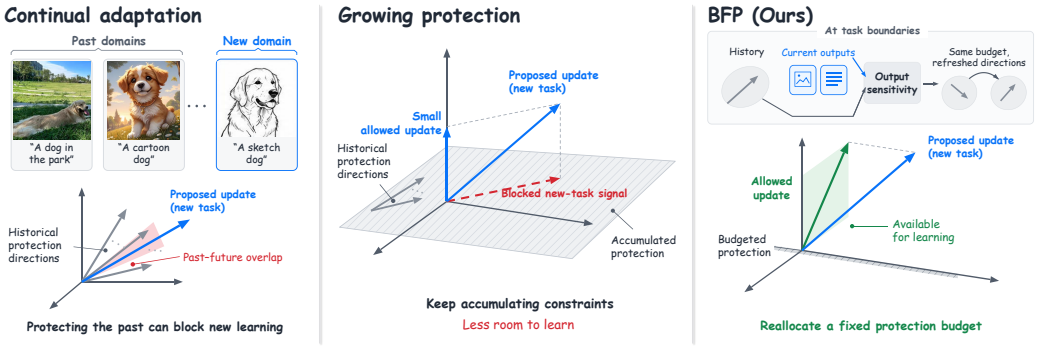


Figure 1: **Growing protection can obstruct new learning. Left:** New tasks can require directions that also matter for past knowledge. **Middle:** Accumulating protection can remove useful components of new-task updates. **Right:** BFP refreshes protected directions under a fixed budget using estimated output sensitivity, leaving the remaining directions available for learning. Geometry is schematic.

according to their estimated influence on model outputs as tasks arrive. The paired diagnostic measures adaptation recoverable from further release.

Our contributions are threefold:

- **We reveal an avoidable cost of historical protection.** Controlled interventions show that accumulating historical constraints can obstruct new-task learning beyond the specified retention requirements. Finite-step theory links recoverable loss to blocked task signal under explicit curvature assumptions.
- **We propose a diagnostic and a budgeted protection method.** The paired PIPL diagnostic isolates the effect of a specified protection release; BFP reduces protection's adaptation cost by refreshing a compact functional protection subspace.
- **BFP improves continual performance and preserves adaptability.** BFP improves the adaptation–retention balance in retrieval and classification. Under the common VLCL budget, it exceeds Pi-CCA by 1.11/1.06 percentage points in image-to-text/text-to-image Recall@1 while reducing pretrained accuracy drop from 3.02 to 1.37 points. On held-out Sketch, it achieves higher adaptation AUC than growing protection with less measured benefit from further constraint release.

2 Related Work

Continual vision–language learning. VLM retention includes cross-modal alignment and zero-shot transfer. Mod-X aligns off-diagonal similarities (Ni et al., 2023); ZSCL uses reference-data distillation and weight ensembling (Zheng et al., 2023). C-CLIP consolidates contrastive knowledge for continual retrieval (Liu et al., 2025); Pi-CCA preserves canonical alignment statistics without replay (Zhang et al., 2026). Following C-CLIP's observation that retention can conceal inadequate learning, we measure protection's adaptation cost alongside historical and pretrained retention.

Subspace protection. OGD protects historical output gradients (Farajtabar et al., 2020); GPM constructs bases from activations (Saha et al., 2021). KeepLoRA excludes principal weight and historical feature directions (Luo et al., 2026), while NuSA-CL uses the current weights' approximate null space (Jo & Kim, 2026). FS-DGPM relaxes protection adaptively (Deng et al., 2021); space decoupling separates stability- and plasticity-related directions (Zhao et al., 2023). BFP reallocates a fixed rank budget using compressed output sensitivity in unchanged parameter coordinates.

Plasticity and capacity budgets. Plasticity can decline without historical constraints (Dohare et al., 2024). PCLR recycles adapter ranks (Meng et al., 2026); QGPM quantizes protection bases and relaxes orthogonality to compensate for quantization error (Kim et al., 2026). Our diagnostic holds the starting model, optimizer, task, and update budget fixed, releases selected directions, and evaluates adaptation with retention to isolate the recoverable effect of that release.

3 PRELIMINARY STUDY

Does historical protection preferentially intercept unseen-task gradients? Across ten runs of CLIP ViT-B/16, *growing protection* retained all realized historical sensitivity directions. After each of the first $t \in \{1, \ldots, 6\}$ domains in a seven-domain stream, we froze the checkpoint and measured raw gradients on 64 next-domain minibatches.

For a task gradient $g_{t,b}$ and an orthonormal historical basis $U_t \in \mathbb{R}^{d \times k_t}$, the intercepted energy fraction is

$$\rho_{t,b} = \frac{\|U_t^\top g_{t,b}\|_2^2}{\|g_{t,b}\|_2^2}. \tag{1}$$

The rank-matched random expectation k_t/d gives excess interception $e_{t,b} = \rho_{t,b} - k_t/d$. Frozen pretrained CLIP features measure minibatch affinity to historical domains (protocol and summaries: Appendix B.4).

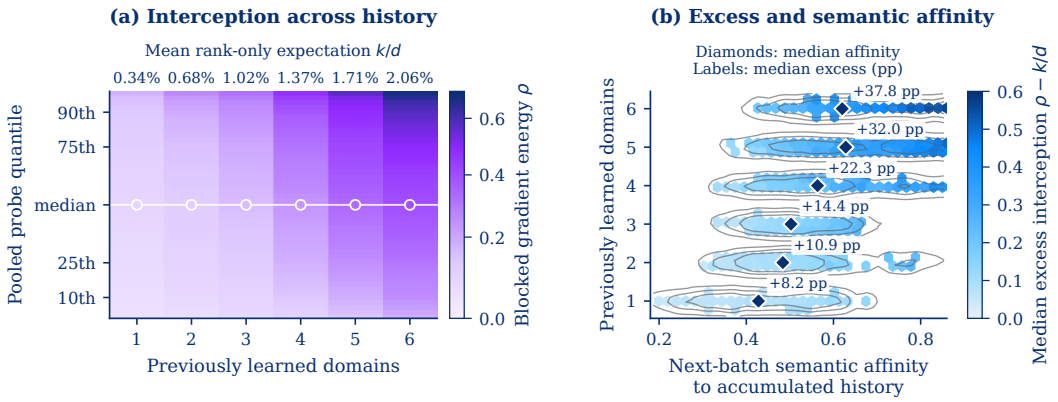


Figure 2: **Historical protection intercepts future-task gradients beyond its rank-only expectation.** **(a)** Interception quantiles; top labels show mean k_t/d. **(b)** Median excess interception versus semantic affinity and history. Each history length pools 640 probes (Appendix B.4.4).

Figure 2 shows that median interception rises from 8.55% at $t = 1$ to 39.90% at $t = 6$, whereas the mean rank-only expectation increases from 0.34% to 2.06%. Higher-affinity minibatches also show greater excess interception at each history length. This overlap motivates the controlled release analysis in Section 4, which jointly evaluates adaptation and historical/pretrained retention.

4 METHOD

4.1 OVERVIEW

The paired PIPL diagnostic evaluates constraint-release gains alongside historical and pretrained retention. BFP projects updates outside at most r leading directions of a compressed output-sensitivity statistic, refreshed by a thin singular value decomposition (SVD) at task boundaries (Figure 3).

4.2 LEARNING SETUP AND PAIRED INTERVENTION

A pretrained vision–language model learns from image–text training sets $\mathcal{D}_1, \ldots, \mathcal{D}_T$ in sequence. We adapt its final image and text projections with rank-eight residual factors, concatenating their trainable parameters into a single vector $\theta \in \mathbb{R}^d$. The factors initially preserve the pretrained function (Appendix A.1). Throughout training, the backbone and temperature are frozen, and the parameterization and joint parameter block stay fixed. Task boundaries are known during training; inference uses one model without task identity. For a minibatch $\mathcal{B} = \{(v_i, c_i)\}_{i=1}^b$, we minimize the symmetric contrastive loss

$$\ell(\theta; \mathcal{B}) = -\frac{1}{2b} \sum_{i=1}^{b} \left[\log \frac{e^{s_{ii}}}{\sum_{j=1}^{b} e^{s_{ij}}} + \log \frac{e^{s_{ii}}}{\sum_{j=1}^{b} e^{s_{ji}}} \right], \qquad s_{ij} = \frac{\phi_\theta(v_i)^\top \psi_\theta(c_j)}{\tau}. \tag{2}$$

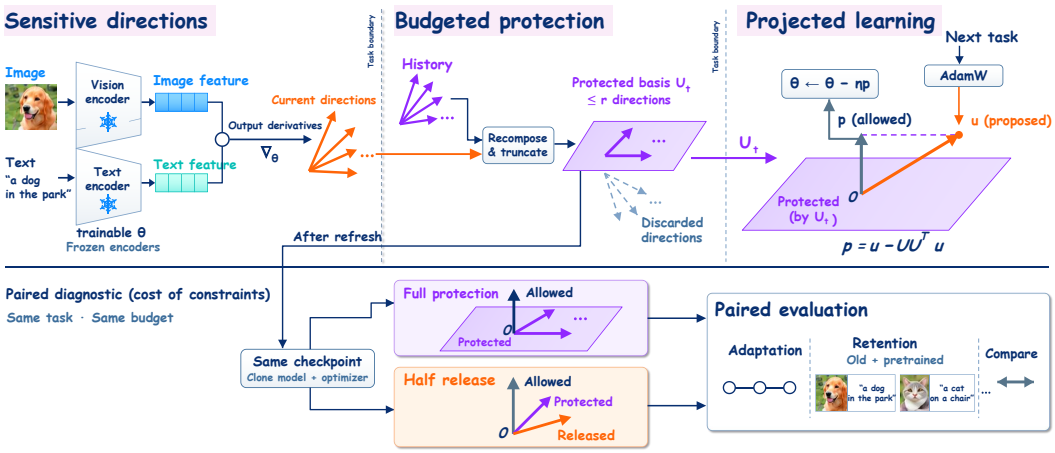


Figure 3: **Budgeted Functional Protection (BFP) and its paired diagnostic. Top:** output-gradient probes of both embeddings identify sensitive projection-adapter directions. At each task boundary, a thin SVD combines the current sketch with the compressed historical statistic and retains at most r leading directions U_t. The next task projects complete AdamW proposals onto the orthogonal complement of U_t. **Bottom:** after the refresh, two branches cloned from the same model and optimizer states learn the same diagnostic task under an identical budget. One keeps full protection; the other retains only the leading half of U_t. We assess adaptation alongside historical-task and pretrained retention. Bases stay fixed, and neither branch alters the parent trajectory. Geometry is schematic.

Here b is the batch size, $\phi_\theta(v), \psi_\theta(c) \in \mathbb{R}^q$ are unit-normalized embeddings, and $\tau > 0$ is the pretrained temperature.

The continual training run, or *parent*, never uses the reserved diagnostic task $\mathcal{D}_\star$. After task t and protection refresh, we clone its model, optimizer, and random states. One branch keeps all k_t columns of U_t; the other keeps its leading $\lfloor k_t/2 \rfloor$ columns, denoted U_t^-. Both use identical diagnostic minibatches for K updates with fixed bases. The release effect is

$$\Delta_t = \mathrm{AUC}_t(U_t^-) - \mathrm{AUC}_t(U_t), \qquad \mathrm{AUC}_t(V) = \frac{1}{K}\int_0^K S_t^V(s)\,\mathrm{d}s. \tag{3}$$

Here $S_t^V(s) \in [0,1]$ is mean bidirectional Recall@1 on a fixed diagnostic test gallery after s updates with basis V. Trapezoidal integration on a fixed grid estimates the normalized area under the curve (AUC). Thus Δ_t includes subsequent optimization effects while holding training history constant.

4.3 A FUNCTIONAL PROTECTION STATISTIC

Output sensitivity captures representation changes even when task-loss gradients are small (Farajtabar et al., 2020). For both modalities at task endpoint θ_t, define

$$z_\theta(x) = \frac{1}{\sqrt{2}}\begin{bmatrix}\phi_\theta(v)\\ \psi_\theta(c)\end{bmatrix}, \qquad J_t(x) = \left.\frac{\partial z_\theta(x)}{\partial \theta}\right|_{\theta_t}, \qquad C_t = \mathbb{E}_{x\sim\mathcal{D}_t}[J_t(x)^\top J_t(x)]. \tag{4}$$

Here $x = (v, c)$, $z_\theta(x) \in \mathbb{R}^{2q}$, $J_t(x) \in \mathbb{R}^{2q\times d}$, and the expectation is uniform over training pairs. For a small displacement δ, the expected squared output change is $\delta^\top C_t \delta + o(\|\delta\|_2^2)$.

We estimate C_t without forming a Jacobian or dense covariance. Sample $m_t = \min(m, |\mathcal{D}_t|)$ pairs uniformly without replacement and independent random-sign vectors $a_i \in \{-1, +1\}^{2q}$, then form the sketch

$$\begin{aligned} h_{t,i} &= J_t(x_i)^\top a_i = \nabla_\theta[a_i^\top z_\theta(x_i)]\big|_{\theta_t}, \\ H_t &= \frac{[h_{t,1}, \ldots, h_{t,m_t}]}{\sqrt{m_t}}, \qquad \bar{H}_t = \frac{H_t}{\|H_t\|_F} \quad \text{if } \|H_t\|_F > 0. \end{aligned} \tag{5}$$

Each column requires one scalar-output backward pass. Conditional on θ_t, $\mathbb{E}[H_t H_t^\top] = C_t$ because $\mathbb{E}[a_i a_i^\top] = I_{2q}$. We skip zero sketches, leaving the protection state unchanged. For nonzero sketches,

normalization gives $\bar{H}_t \bar{H}_t^\top$ unit trace. Its conditional expectation generally differs from $C_t / \operatorname{tr}(C_t)$ (Appendix A.2).

4.4 Reallocating a fixed protection budget

With a budget of $r < d$ protected directions, we minimize the worst-case sensitivity in the remaining *free space*, where updates are allowed. For a positive semidefinite statistic C, this gives

$$\min_{\substack{U \in \mathbb{R}^{d \times k},\, U^\top U = I_k \\ 0 \le k \le r}} \; \max_{\substack{\|w\|_2 = 1 \\ U^\top w = 0}} w^\top C w = \lambda_{r+1}(C), \qquad P_U = I_d - U U^\top. \tag{6}$$

Here w is an allowed unit displacement, P_U projects onto the free space, and $\lambda_j(C)$ is the jth largest eigenvalue. Leading $\min(r, \operatorname{rank} C)$ eigenvectors attain the minimum; intersection with the leading $(r+1)$-dimensional eigenspace gives the lower bound (Appendix D.1).

BFP stores a compressed statistic $\widetilde{C} = U \Lambda U^\top$ using orthonormal U, nonnegative diagonal Λ, and a count n of nonzero sketches. Starting with an empty basis and $n = 0$, each nonzero task sketch updates the state by

$$\begin{aligned} X &= \left[\sqrt{\frac{n}{n+1}}\, U \Lambda^{1/2},\; \frac{\bar{H}_t}{\sqrt{n+1}} \right] = Q \Sigma V^\top, \\ U^+ &= Q_{:,1:k}, \qquad \Lambda^+ = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_k^2), \qquad k = \min(r, \operatorname{rank} X). \end{aligned} \tag{7}$$

The thin SVD orders σ_j decreasingly. We replace (U, Λ, n) with $(U^+, \Lambda^+, n+1)$ without renormalizing Λ^+. The basis solves (6) for the *available* statistic $X X^\top$. Task t fixes U_{t-1}, then refreshes to U_t. The joint block uses $r = 16$, distinct from adapter rank, and $m = 64$.

Repeated truncation can permanently discard a direction that later dominates the full historical average, even without sampling error or Jacobian drift. BFP optimizes its compressed statistic; Appendix A.3 gives a two-dimensional counterexample and an exact identity quantifying residual sensitivity relative to the full realized history.

4.5 Projected learning and controlled diagnosis

During task t, we project each optimizer proposal onto the free space, making the smallest Euclidean change needed to satisfy the protection constraints:

$$p = \operatorname*{arg\,min}_{v:\, U_{t-1}^\top v = 0} \frac{1}{2} \|v - u\|_2^2 = P_{U_{t-1}} u, \qquad \theta \leftarrow \theta - \eta p. \tag{8}$$

Here $u \in \mathbb{R}^d$ is the proposal, p its projection, and $\eta > 0$ the learning rate. AdamW computes moments from raw gradients, applying momentum, coordinatewise preconditioning, and weight decay *before* projection. Projection enforces constraints; descent depends on gradient alignment (Appendix D.1). Appendix A.1 gives the algorithm and numerical conventions. Storage and each projection cost $O(dr)$; a boundary requires m_t output backward passes and $O(d(r + m_t)^2)$ SVD work.

At a shared state, release from U to a nested basis U^- restores the gradient component along the released directions:

$$\|P_{U^-} g\|_2^2 - \|P_U g\|_2^2 = \|(U U^\top - U^- (U^-)^\top) g\|_2^2 \ge 0, \qquad g = \nabla_\theta \ell(\theta; \mathcal{B}). \tag{9}$$

An additional half-release branch controls step length by matching the full branch's actual displacement norm at each step, with directions computed from its own evolving state. Appendix B.2 specifies this control, zero-proposal failures, and equal-rank orientation controls.

We diagnose BFP and *growing protection*, which retains every nonzero singular direction in (7), using shared initialization, task order, data schedule, and optimizer settings. Parents train on seven VLCL domains, reserving Sketch (Liu et al., 2025) as $\mathcal{D}_\star$. At $t \in \{1, 4, 7\}$, branches share diagnostic minibatches, $K = 200$ updates, and the evaluation grid. The contrast $\Delta_7 - \Delta_1$ tracks release effects as both history length and learned state evolve.

We classify protection as *excessive* only when release improves adaptation while satisfying retention noninferiority. For each parent at $t = 7$, we require

$$L(\Delta_7) > 0.01, \qquad L(R_7^- - R_7) \geq -0.01, \qquad L(Z_7^- - Z_7) \geq -0.01. \tag{10}$$

Here R_7 is mean bidirectional Recall@1 across the seven historical domains, and Z_7 is macro-averaged zero-shot classification accuracy on a fixed six-dataset suite. Superscript $-$ denotes half release; scores lie in $[0, 1]$, so each margin is one percentage point. The operator L gives a multiplicity-adjusted one-sided lower confidence bound over ten independent paired runs. Appendix B.3 specifies the metrics, bounds, and primary hypothesis family. The half-release rule is fixed across runs and checkpoints.

4.6 Finite-step protection-induced plasticity loss

Consider the quadratic loss $f(\delta) = f_0 - b^\top \delta + \frac{1}{2}\delta^\top \mathsf{H}\delta$, where $\delta = \theta - \theta_0$ and $b = -\nabla f(0)$. Let V span the original free space $\ker U^\top$ and W the newly released directions, so $Z = [V, W]$ spans $\ker(U^-)^\top$. Both blocks are nonempty, Z has orthonormal columns, and $0 < \mu I \preceq Z^\top \mathsf{H} Z \preceq LI$. Starting at $\delta_0^Q = 0$, both branches use fixed spaces and a shared step size $0 < \eta \leq 1/L$. Their projected gradient updates and the loss recovered after K steps are

$$\delta_{k+1}^Q = \delta_k^Q - \eta Q Q^\top \nabla f(\delta_k^Q), \quad Q \in \{V, Z\}, \qquad \mathcal{G}_K = f(\delta_K^V) - f(\delta_K^Z). \tag{11}$$

Theorem 1 (Finite-step recovery from released task signal)**.** *Suppose $V^\top \mathsf{H} W = 0$, so curvature does not couple the original free and newly released directions. Let $c = W^\top b$ be the released task signal and $D = W^\top \mathsf{H} W$ its curvature matrix. Write $\mu_D = \lambda_{\min}(D)$, $L_D = \lambda_{\max}(D)$, and $\Gamma = \frac{1}{2} c^\top D^{-1} c$. For every integer $K \geq 1$,*

$$\begin{gathered} \mathcal{G}_K = \tfrac{1}{2} c^\top D^{-1} [I - (I - \eta D)^{2K}] c, \\ [1 - (1 - \eta\mu_D)^{2K}]\Gamma \leq \mathcal{G}_K \leq [1 - (1 - \eta L_D)^{2K}]\Gamma. \end{gathered} \tag{12}$$

Both bounds are attainable. Moreover, $\mathcal{G}_K > 0$ exactly when $c \neq 0$, and $\mathcal{G}_K \to \Gamma$ as $K \to \infty$.

Proof sketch. The original free coordinates evolve identically, so their loss contributions cancel. Solving along W gives (12), whose eigenvalue factors yield the bounds. Equality holds when c aligns with an extremal eigenvector of D (full proof in Appendix D.7).

Curvature coupling can reverse finite-step loss ordering; Appendix D.8 bounds its effect and Appendix D.11 gives exact counterexamples. The theory covers deterministic projected gradient descent; AdamW dynamics, Recall@1 gains, and retention require separate evaluation. Appendix D gives general decompositions and a smooth-loss extension under explicit conditions.

5 Experiments

We examine joint learning and retention (Q1, Section 5.2), whether protection obstructs adaptation (Q2, Section 5.3), the value of individual directions (Q3, Section 5.4), and robustness and cost (Q4, Section 5.5).

5.1 Experimental setup

Benchmarks and metrics. We evaluate VLCL retrieval (Liu et al., 2025), full-/five-shot MTIL classification (Zheng et al., 2023; Jo & Kim, 2026), and task-agnostic X-TAIL (Zhang et al., 2026). Metrics comprise bidirectional Recall@1, pretrained accuracy drop, and classification Transfer/Average/Last. CIFAR-100 class increments and 32-stage VLCL test long-sequence behavior. PIPL is measured by paired adaptation AUC gains on held-out Sketch, jointly with historical and pretrained retention. Splits, metric definitions, and statistical tests appear in Appendix B.1.

Comparison protocol. We compare with C-CLIP (Liu et al., 2025), Pi-CCA (Zhang et al., 2026), KeepLoRA (Luo et al., 2026), NuSA-CL (Jo & Kim, 2026), and established baselines listed in Appendix B.1.3. Matched comparisons share pretrained checkpoints, splits, and evaluation settings,

with source schedules separated from common-update budgets. External methods retain their architectures and auxiliary resources. Fixed-coordinate controls isolate protection capacity, orientation, and optimizer-state effects (Appendix C.4).

Backbones and implementation. We use OpenAI CLIP ViT-B/16, ViT-B/32, and ViT-L/14 (Radford et al., 2021), with ViT-B/16 as the primary backbone. BFP adapts the two final projections at rank eight, using protection rank $r = 16$, $m = 64$ probes, and AdamW. Training budgets, optimizer settings, hardware, and protection-state measurements are detailed in Appendix B.1.4.

5.2 Q1: Can BFP improve learning and retention together?

Continual retrieval and pretrained retention. Under the common budget, BFP achieves 42.05/39.58 I2T/T2I $R@1$, exceeding Pi-CCA by 1.11/1.06 percentage points (pp) while reducing PD_6 from 3.02 to 1.37 pp (Table 1). BFP also improves over Pi-CCA under the source protocol; domain-level results show gains over C-CLIP in every retrieval domain and direction (Table 6).

Continual classification. BFP reaches full-shot Last scores of 86.7 and 87.0 under the NuSA-CL and KeepLoRA protocols, respectively (Table 2). Under the common budget, its Last scores lead the compared methods by 0.7/1.1 pp on full-/five-shot MTIL and 1.5 pp on task-agnostic X-TAIL (Table 8). The gains therefore extend across retrieval, classification, and evaluation with a shared vocabulary.

Table 1: **VLCL retrieval and pretrained retention.** ViT-B/16; $R@1$ (%, ↑) and PD (pp, ↓).

	Source protocol				Common budget		
Method	I2T	T2I	PD_{I1K}	PD_{C100}	I2T	T2I	PD_6
Frozen CLIP (Radford et al., 2021)	8.79	20.86	—	—	8.79	20.86	0.00
Sequential LoRA (Hu et al., 2022)	—	—	—	—	33.12	32.46	10.48
EWC (Kirkpatrick et al., 2017)	27.29	26.49	26.97	24.08	30.11	29.44	11.75
ZSCL (Zheng et al., 2023)	31.25	30.83	18.13	14.68	35.28	34.41	5.82
Mod-X (Ni et al., 2023)	28.18	27.36	23.52	22.53	31.67	30.88	9.64
MoE-Adapters (Yu et al., 2024)	29.66	30.03	20.68	16.90	34.04	33.76	7.11
DKR (Cui et al., 2024)	28.43	28.93	21.85	20.42	32.57	32.88	8.36
C-CLIP (Liu et al., 2025)	40.83	37.97	7.42	5.29	40.11	37.74	4.21
Pi-CCA (Zhang et al., 2026)	41.57	38.84	5.96	4.37	40.94	38.52	3.02
BFP (ours)	**42.36**	**39.35**	**4.74**	**3.61**	**42.05**	**39.58**	**1.37**

Table 2: **MTIL classification.** ViT-B/16; all metrics in percent (↑).

	Full-shot			Five-shot		
Method	Transfer	Average	Last	Transfer	Average	Last
(a) NuSA-CL protocol						
Full fine-tuning (Jo & Kim, 2026)	44.6	55.9	77.3	—	—	—
LoRA (Hu et al., 2022)	63.9	70.1	79.9	60.4	66.8	72.2
InfLoRA (Liang & Li, 2024)	66.2	74.2	83.6	66.8	68.9	74.8
NuSA-CL (Jo & Kim, 2026)	68.6	75.1	82.8	68.1	70.3	75.4
BFP (ours)	**70.1**$_{+1.5}$	**78.2**$_{+3.1}$	**86.7**$_{+3.1}$	**70.1**$_{+2.0}$	**71.6**$_{+1.3}$	**77.0**$_{+1.6}$
(b) KeepLoRA-based protocol						
ZSCL (Zheng et al., 2023)	68.1	75.4	83.6	65.7	69.1	74.3
MoE-Adapters (Yu et al., 2024)	68.9	76.7	85.0	66.5	70.2	75.4
InfLoRA (Liang & Li, 2024)	67.4	76.3	85.2	66.1	69.7	75.2
SD-LoRA (Wu et al., 2025)	67.1	72.5	79.2	64.8	67.6	72.9
KeepLoRA (Luo et al., 2026)	69.0	77.5	86.1	67.6	71.3	76.4
BFP (ours)	**70.2**$_{+1.2}$	**78.4**$_{+0.9}$	**87.0**$_{+0.9}$	**68.8**$_{+1.2}$	**72.2**$_{+0.9}$	**77.3**$_{+0.9}$

Takeaway (Q1). Selective protection improves the observed learning–retention balance.

5.3 Q2: Does historical protection itself obstruct adaptation?

Paired release and history dependence. Growing protection satisfies the PIPL criterion: $L(\Delta_7) = 5.70$ pp, with both retention bounds above -1 pp (Table 3). Its release effect rises from 1.25 to 6.75 pp, with a 4.33 pp lower bound on this increase, as both history length and learned state change (early/middle scores: Table 12). BFP has higher unreleased AUC (57.86 versus 50.42) and a smaller mean release effect (0.76 versus 6.75 pp), below the excessive-protection threshold. This supports reduced observed obstruction; reporting conventions appear in Appendix B.5.

Table 3: **Paired adaptation and retention on held-out Sketch.** Scores are percentages; effects and bounds are in percentage points.

(a) Absolute performance ($t = 7$)

Parent	AUC_7	AUC_7^-	R_7	R_7^-	Z_7	Z_7^-
Growing protection	50.42 ±1.73	57.17 ±1.84	**51.36** ±0.82	**51.04** ±0.84	63.12 ±0.44	62.98 ±0.47
BFP (ours)	**57.86** ±1.55	**58.62** ±1.59	50.98 ±0.78	50.81 ±0.79	**63.34** ±0.41	**63.25** ±0.43

(b) Release effects and lower bounds

Parent	Δ_1	Δ_4	Δ_7	$L(\Delta_7)$	$L(dR_7)$	$L(dZ_7)$
Growing protection	1.25	3.96	6.75	5.70	−0.68	−0.41
BFP (ours)	0.44	0.62	0.76	0.33	−0.48	−0.33

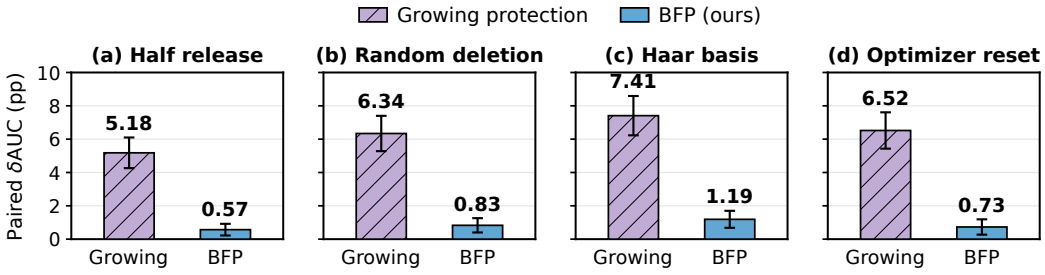


Figure 4: **Release effects persist under mechanism controls.** Paired Sketch AUC changes at $t = 7$ (ten runs; mean ± sample SD). Panels (a–c) replay full-branch displacement norms; (d) compares reset-half with reset-full without rescaling. Full results: Table 13.

Step length, optimizer state, and orientation. Growing protection retains a 5.18 pp half-release AUC gain under norm replay and a 6.52 pp gain after resetting both optimizer branches; BFP's corresponding gains are 0.57/0.73 pp (Figure 4). The obstruction therefore extends beyond update length and inherited optimizer state. Random deletion and Haar replacement produce larger AUC gains at a greater retention cost (Table 13).

Takeaway (Q2). Paired release tests reveal lost learning capacity concealed by retained-task scores.

5.4 Q3: Which protected directions are worth keeping?

Capacity versus direction choice. With identical adapters and schedules (Table 11), rank-16 BFP retains more historical/pretrained performance but has lower adaptation AUC than frozen first-task or fixed random bases. Removing protection gives the highest AUC and lowest retention. Refreshing directions at fixed rank thus shifts the balance toward retention.

Marginal utility of individual directions. We release one protected direction at a time and repeat 200 Sketch updates from the same parent checkpoint at $t = 7$ (Appendix B.6). Figure 5 maps the resulting adaptation gain G_i against retention damage D_i.

Under growing protection, 16.4% of directions fall in the high-gain, low-damage region and account for 47.3% of summed positive single-direction AUC gains; only 5.0% of BFP directions fall there. Growing protection's release gains correlate strongly with gradient overlap q_i (Pearson $r = 0.883$),

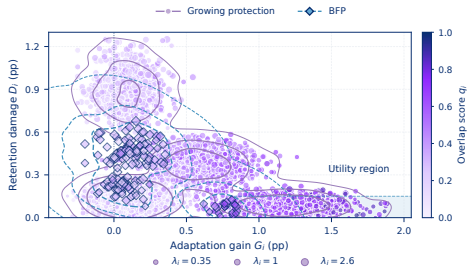


Figure 5: **Protection utility by direction.** Growing protection: circles/solid contours; BFP: diamonds/dashed contours. Color shows overlap q_i; area increases with $\log(1 + \lambda_i)$. Shading marks high adaptation gain ($G_i > 0.8$ pp), low retention damage ($D_i < 0.15$ pp).

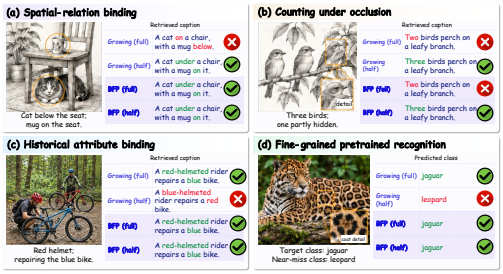


Figure 6: **Adaptation and retention cases.** After 200 Sketch updates ($t = 7$): (a–b) compositional retrieval, (c) historical attribute binding, (d) pretrained fine-grained recognition. Rows pair full/half protection within each parent. Amber: visual cues; green checks/red crosses: correct/incorrect predictions.

but weakly with stored sensitivity $\log(1 + \lambda_i)$ ($r = 0.047$). Obstruction is therefore concentrated in directions with high overlap but little measured retention utility. Each gain is measured by releasing one direction while keeping all others protected.

Behavioral consequences. Figure 6 shows spatial-relation and occluded-counting errors on Sketch, including a residual BFP error under full protection. Release loses historical attribute binding and pretrained fine-grained recognition under growing protection, while BFP preserves both predictions. These within-parent comparisons expose release's behavioral gains and retention costs.

Takeaway (Q3). A direction's value depends on both preserved behavior and obstructed learning.

5.5 Q4: How robust and costly is bounded protection?

Backbones and longer streams. BFP leads both retrieval directions across three backbones and both 32-stage variants (Table 9). At fixed total updates, moving from eight to 32 interleaved stages reduces ViT-B/16 I2T/T2I by 0.59/0.52 pp for BFP versus 1.23/1.18 pp for Pi-CCA. BFP also leads CIFAR-100 at 10, 20, and 50 stages under native and common budgets, with a common-budget Last advantage of 1.21–1.51 pp over the strongest baseline (Table 7).

Protection overhead. Along the diagnostic stream, BFP keeps rank 16 and 2.250 MiB of protection state; growing protection reaches rank 438–447 and 62.31 MiB at $t = 7$ (Table 14). Late projection/refresh costs are 0.044 ms/1.94 s versus 0.123 ms/4.08 s, respectively. These timings exclude other training costs. More task boundaries add probing and SVD work at a fixed update budget.

Takeaway (Q4). BFP bounds protection storage; refresh cost depends on the number of task boundaries.

6 Conclusion

We studied the *stability trap*: the avoidable adaptation cost of historical protection in continual learning of multimodal foundation models. Focusing on vision–language models, we combined a paired diagnostic of protection-induced plasticity loss with a finite-step analysis of constraint release. Budgeted Functional Protection (BFP) reduces this cost by continually reallocating a fixed protection budget according to output sensitivity. Experiments show that accumulated constraints can obstruct adaptation beyond the prescribed retention requirements, while BFP improves retrieval and classification performance and leaves less adaptation gain recoverable by release. The learning cost of protection thus provides a criterion for designing and evaluating continual learners. An open question is how to adapt protection budgets to changing task relationships while sustaining adaptation and retention over long streams.

AI USE STATEMENT

AI tools were used to assist in organizing and interpreting experimental results, polishing the manuscript's language, and formatting the manuscript. The authors take full responsibility for all text and results presented in this paper.

REFERENCES

Mehdi Abbana Bennani, Thang Doan, and Masashi Sugiyama. Generalisation guarantees for continual learning with orthogonal gradient descent. *arXiv preprint arXiv:2006.11942*, 2020. URL https://arxiv.org/abs/2006.11942.

Cheng Chen, Junchen Zhu, Xu Luo, Heng Tao Shen, Jingkuan Song, and Lianli Gao. CoIN: A benchmark of continual instruction tuning for multimodel large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1844.

Jinpeng Chen, Runmin Cong, Yuzhi Zhao, Hongzheng Yang, Guangneng Hu, Horace Ip, and Sam Kwong. SEFE: Superficial and essential forgetting eliminator for multimodal continual instruction tuning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 7982–8001. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25n.html.

Zhenyu Cui, Yuxin Peng, Xun Wang, Manyu Zhu, and Jiahuan Zhou. Continual vision-language retrieval via dynamic knowledge rectification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11704–11712, 2024. doi: 10.1609/aaai.v38i10.29054. URL https://ojs.aaai.org/index.php/AAAI/article/view/29054.

Danruo Deng, Guangyong Chen, Jianye Hao, Qiong Wang, and Pheng-Ann Heng. Flattening sharpness for dynamic gradient projection memory benefits continual learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9b16759a62899465ab21e2e79d2ef75c-Abstract.html.

Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632: 768–774, 2024. doi: 10.1038/s41586-024-07711-7. URL https://www.nature.com/articles/s41586-024-07711-7.

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. PMLR, 2020. URL https://proceedings.mlr.press/v108/farajtabar20a.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Yujin Jo and Taesup Kim. Memory-free continual learning with null space adaptation for zero-shot vision-language models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.21175.

Borui Kang, Jinrui Gu, Tao Feng, Qi Fan, Yinghuan Shi, Lei Wang, Wenbin Li, and Yang Gao. Don't forget why you started: Tackling dual forgetting in vision-language continual learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Dongjun Kim, Seohyeon Cha, Huancheng Chen, Chaining Wang, and Haris Vikalo. Quantized gradient projection for memory-efficient continual learning. In *International Conference on Learning Representations*, pp. 21920–21945, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/24e8b46430df965674221665816a4964-Abstract-Conference.html.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://arxiv.org/abs/1612.00796.

Yan-Shuo Liang and Wu-Jun Li. InfLoRA: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2404.00228.

Jiahong Liu, Ming Shen, Xiaohao Liu, Rex Ying, Menglin Yang, Tat-Seng Chua, and Irwin King. Hyperbolic multimodal continual learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2608.09572.

Wenzhuo Liu, Fei Zhu, Longhui Wei, and Qi Tian. C-CLIP: Multimodal continual learning for vision-language model. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/72fb9ab442fc60b7ae5d53bf6b478273-Abstract-Conference.html.

Mao-Lin Luo, Zi-Hao Zhou, Yi-Lin Zhang, Yuanyu Wan, Min-Ling Zhang, and Tong Wei. KeepLoRA: Continual learning with residual gradient adaptation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/dc3ca8bcd613e43ce540352b58d55d6d-Abstract-Conference.html.

Weicheng Meng, Jingyang Qiao, Zhizhong Zhang, Shaohui Liu, and Yuan Xie. PCLR: Progressively compressed LoRA for multimodal continual instruction tuning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/5a5acfd0876c940d81619c1dc60e7748-Abstract-Conference.html.

Zixuan Ni, Longhui Wei, Siliang Tang, Yueting Zhuang, and Qi Tian. Continual vision-language representation learning with off-diagonal information. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26129–26149. PMLR, 2023. URL https://proceedings.mlr.press/v202/ni23c.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. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2103.09762.

Yichen Wu, Hongming Piao, Long-Kai Huang, Renzhen Wang, Wanhua Li, Hanspeter Pfister, Deyu Meng, Kede Ma, and Ying Wei. SD-LoRA: Scalable decoupled low-rank adaptation for class incremental learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/92f43b1d33fae4aa1958f75317f0cec1-Abstract-Conference.html.

Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23219–23230, 2024. doi: 10.1109/CVPR52733.2024.02191. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yu_Boosting_Continual_Learning_of_Vision-Language_Models_via_Mixture-of-Experts_Adapters_CVPR_2024_paper.html.

Jiayu Zhang, Chuangxin Zhao, Canran Xiao, Ruibo Duan, Wenyi Mo, Haoyu Gao, and Wenshuo Wang. Pi-CCA: Prompt-invariant CCA certificates for replay-free continual multimodal learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e4165c96702bac5f4962b70f3cf2f136-Abstract-Conference.html.

Zhen Zhao, Zhizhong Zhang, Xin Tan, Jun Liu, Yanyun Qu, Yuan Xie, and Lizhuang Ma. Rethinking gradient projection continual learning: Stability / plasticity feature space decoupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3718–3727, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Zhao_Rethinking_Gradient_Projection_Continual_Learning_Stability__Plasticity_Feature_Space_CVPR_2023_paper.html.

Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, and Yang You. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19125–19136, 2023. doi: 10.1109/ICCV51070.2023.01752. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zheng_Preventing_Zero-Shot_Transfer_Degradation_in_Continual_Learning_of_Vision-Language_Models_ICCV_2023_paper.html.

Appendix Contents

A Technical Details 14
- A.1 Complete algorithm and fixed implementation 14
- A.2 Probe estimation and the nonzero event 15
- A.3 Compression error and failure case 15
- A.4 Scope and relation to prior methods 16

B Experimental Details 17
- B.1 Benchmarks and training configuration 17
- B.2 Deterministic intervention protocol 20
- B.3 Metrics, uncertainty, and decision rules 21
- B.4 Preliminary-study protocol 22
- B.5 Source protocols and reporting conventions 23
- B.6 Direction-level protection-utility protocol 24

C Additional Experimental Results 25
- C.1 Preliminary-study numerical summaries 25
- C.2 Benchmark comparisons 25
- C.3 Direction-level protection-utility statistics 28
- C.4 Mechanism controls and protection overhead 29

D Proofs 31
- D.1 Spectral and projection geometry 31
- D.2 Accumulated compression error 31
- D.3 Finite-step model and notation 31
- D.4 Restricted geometry and elimination of free coordinates 32
- D.5 Projected dynamics and sharp spectral residuals 33
- D.6 Exact finite-step decomposition and sufficient budget 34
- D.7 Proof of the main finite-step recovery theorem 35
- D.8 Curvature coupling and second-order correction 36
- D.9 Loss-area identities on a fixed evaluation grid 39
- D.10 Extension to smooth, nonquadratic losses 40
- D.11 Counterexamples to unconditional finite-step improvement 43
- D.12 Degenerate spaces, coordinate invariance, and scope 44

A TECHNICAL DETAILS

A.1 COMPLETE ALGORITHM AND FIXED IMPLEMENTATION

A.1.1 PARAMETERIZATION AND CONFIGURATION

The backbone is the public OpenAI CLIP ViT-B/16 checkpoint (Radford et al., 2021). Only its two final embedding projections are adapted. In column-vector convention, for modality $\nu \in \{\mathrm{img}, \mathrm{txt}\}$,

$$W_\nu = W_\nu^0 + B_\nu A_\nu, \qquad A_\nu \in \mathbb{R}^{8\times p_\nu}, \quad B_\nu \in \mathbb{R}^{512\times 8}, \qquad \theta = \mathrm{vec}(A_{\mathrm{img}}, B_{\mathrm{img}}, A_{\mathrm{txt}}, B_{\mathrm{txt}}). \tag{13}$$

Here W_ν^0 is frozen, $p_{\mathrm{img}} = 768$, and $p_{\mathrm{txt}} = 512$, giving $d = 18{,}432$. Entries of A_ν are independently initialized from $\mathcal{N}(0, 1/p_\nu)$ and $B_\nu = 0$; the residual scale is one. The embeddings are normalized after applying these projections. The four column-wise vectorizations form one joint protection block in the displayed order. The factor coordinates and scale remain fixed throughout the stream.

Table 4: Fixed primary configuration. Diagnostic thresholds apply only to evaluation and are specified in Appendix B.3.

Quantity	Specification
Adaptation	Two final projection adapters, rank 8, one joint block
Protection	$r = 16$; $m = 64$ output probes per boundary
Optimizer	AdamW; $\eta = 10^{-4}$; $(\beta_1, \beta_2) = (0.9, 0.999)$; $\epsilon_{\mathrm{opt}} = 10^{-8}$; weight decay 0.01
Training budget	1,000 updates per parent domain; batch size 256; constant learning rate; no gradient clipping
Input function	CLIP resize and center crop to 224×224, published normalization and tokenizer; no stochastic augmentation or dropout
Precision	FP32 model and gradients; FP64 sketch accumulation, SVD, and projection; cast the final projected update to parameter precision
Selection rule	Fixed protection rank 16 and probe budget 64, with no test-based selection

A.1.2 UPDATE ORDERING AND NUMERICAL RULES

For global optimizer step s, use $g_s = \nabla_\theta \ell(\theta_{s-1}; \mathcal{B}_s)$. AdamW updates its moments by $m_s = \beta_1 m_{s-1} + (1-\beta_1) g_s$ and $v_s = \beta_2 v_{s-1} + (1-\beta_2) g_s^{\odot 2}$. Bias correction gives $\widehat{m}_s = m_s/(1-\beta_1^s)$ and $\widehat{v}_s = v_s/(1-\beta_2^s)$. The complete proposal is $u_s = \widehat{m}_s \oslash (\sqrt{\widehat{v}_s} + \epsilon_{\mathrm{opt}}) + 0.01\theta_{s-1}$, where $\odot$ and $\oslash$ denote elementwise operations. Apply (8) once, using a single learning rate for all coordinates. Moments and the step counter persist across parent tasks. Boundary probing does not update these states.

A sketch is skipped exactly when all its computed entries are zero. Its Frobenius norm is evaluated with scaled accumulation to avoid underflow; no additional small-norm cutoff is applied. For an SVD with largest singular value σ_1 and w input columns, numerical rank counts $\sigma_j > \epsilon_{32} \max(d, w)\sigma_1$, with $\epsilon_{32} = 2^{-23}$. The retained budget is the smaller of this rank and r. This cutoff is fixed by FP32 probe precision, independently of evaluation scores. Singular-vector signs are fixed by making the first largest-magnitude coordinate positive. Within an exactly repeated positive singular value, use coordinate-ordered Gram–Schmidt on projected coordinate vectors to fix the basis before selecting columns. Nonfinite quantities and empty tasks terminate the run and are recorded as failures.

Algorithm 1 stores $O(dr)$ protection entries and uses $O(d(r + m_t))$ temporary entries at a boundary. It never constructs a dense $d \times d$ matrix. Growing protection uses the same algorithm with the rank cap removed; its storage grows with the realized probe span. It is untruncated in exact arithmetic, while the shared numerical-rank rule still applies in computation. Its realized rank is recorded at every boundary and summarized at $t = 1, 4, 7$ in Table 14. With seven domains and $m = 64$, its rank is at most 448, below d.

Algorithm 1 BFP on a fixed joint parameter vector

Require: Parameters θ_0; nonempty task sets $\mathcal{D}_{1:T}$; rank r; probe budget m

$\theta \leftarrow \theta_0$; $(U, \Lambda, n) \leftarrow ([\,], [\,], 0)$; initialize AdamW
for $t = 1, \ldots, T$ **do**
  for each of the 1,000 prescribed minibatches $\mathcal{B}_{t,s}$ **do**
    Compute raw gradient g of (2); update AdamW moments
    Form the complete proposal u, including weight decay
    $\theta \leftarrow \theta - \eta[u - U(U^\top u)]$
  end for
  $\theta_t \leftarrow \theta$; $m_t \leftarrow \min(m, |\mathcal{D}_t|)$
  Sample m_t training pairs without replacement and independent signs a_i
  Form H_t by (5)
  if $H_t \neq 0$ **then**
    Normalize H_t; compute the thin SVD in (7)
    Retain at most r directions using Appendix A.1.2
    $(U, \Lambda, n) \leftarrow (U^+, \Lambda^+, n + 1)$
  end if
  Save the boundary checkpoint; discard probe examples and temporary matrices
end for
return θ_T; inference does not use the protection statistics

A.2 Probe estimation and the nonzero event

For the finite training set, differentiability gives $z_{\theta_t+\delta}(x) - z_{\theta_t}(x) = J_t(x)\delta + o(\|\delta\|_2)$. Squaring and averaging yields the local expansion following (4). At fixed x, the sign identity $\mathbb{E}_a[aa^\top] = I$ gives $\mathbb{E}_a[(J_t^\top a)(J_t^\top a)^\top] = J_t^\top J_t$. Uniform sampling without replacement preserves the sample-mean expectation, proving the unconditional identity $\mathbb{E}[H_t H_t^\top \mid \theta_t] = C_t$.

Let $\mathcal{V}_t = \{\|H_t\|_F > 0\}$. If $\Pr(\mathcal{V}_t \mid \theta_t) > 0$, normalization on this event satisfies

$$S_t = \frac{H_t H_t^\top}{\operatorname{tr}(H_t H_t^\top)} \quad \text{on } \mathcal{V}_t, \qquad \mathbb{E}[S_t \mid \theta_t, \mathcal{V}_t] \neq \frac{C_t}{\operatorname{tr}(C_t)} \quad \text{in general.} \tag{14}$$

Conditioning and self-normalization both change the estimand, so equality depends on the probe distribution. A zero realization can occur with nonzero C_t: for one sample with Jacobian $J = (1, 1, 0, 0)^\top$, the probe is $a_1 + a_2$, which vanishes with probability $1/2$, although $J^\top J = 2$. If $C_t = 0$, all probes are zero almost surely and the state remains unchanged. Unit-trace weighting applies to updates with nonzero sketches.

Preserving both embeddings also controls their pairwise cosine score. For unit embeddings (ϕ, ψ) before an update and (ϕ', ψ') afterward, $|\phi'^\top \psi' - \phi^\top \psi| \leq \|\phi' - \phi\|_2 + \|\psi' - \psi\|_2 \leq 2\|z' - z\|_2$. This sufficient criterion can overprotect: a common orthogonal rotation preserves cosine scores while changing the embeddings. We therefore evaluate local sensitivity alongside measured retention.

A.3 Compression error and failure case

A.3.1 Stored statistic and full-history reference

Only (U, Λ, n) is stored by BFP. For analysis, index the valid sketches by $j = 1, \ldots, n$, write $S_j = H_j H_j^\top$, and define the full realized mean $C_n^{\text{all}} = n^{-1} \sum_{j=1}^n S_j$. This full-history mean serves as an analytical reference for the stored statistic. Let $\widetilde{C}_0 = 0$ and

$$M_j = \frac{j-1}{j}\widetilde{C}_{j-1} + \frac{1}{j}S_j, \qquad \widetilde{C}_j = \mathcal{T}_r(M_j), \qquad E_j = M_j - \widetilde{C}_j. \tag{15}$$

The operator $\mathcal{T}_r$ retains at most the leading r nonzero eigenpairs; $M_j = X_j X_j^\top$ is the available matrix represented by the thin factor, and $E_j \succeq 0$ is its discarded residual.

Proposition 2 (Accumulated discarded sensitivity)**.** *In exact arithmetic,* $D_n = C_n^{\mathrm{all}} - \widetilde{C}_n$ *satisfies*

$$\begin{aligned} D_n &= \frac{1}{n}\sum_{j=1}^{n} jE_j \succeq 0, \qquad \operatorname{tr}(D_n) = 1 - \operatorname{tr}(\widetilde{C}_n), \\ \|D_n\|_2 &\le \frac{1}{n}\sum_{j=1}^{n} j\lambda_{r+1}(M_j). \end{aligned} \tag{16}$$

The proof is given in Appendix D.2. Numerical rank truncation adds discarded eigenpairs to E_j. The first two identities still hold for the idealized matrices with that truncation, while the bound uses $\|E_j\|_2$ in place of $\lambda_{r+1}(M_j)$; floating-point roundoff contributes separately.

For $U_n^\top p = 0$, $p^\top C_n^{\mathrm{all}} p = p^\top D_n p \le \|D_n\|_2 \|p\|_2^2$. The residual sensitivity is thus controlled by the accumulated compression error $\|D_n\|_2$. To distinguish other errors, let $G_j(\theta)$ be the population output covariance for the training distribution at valid boundary j, and let ϑ_j be that boundary's parameters. Set $A_n = n^{-1}\sum_j G_j(\vartheta_j)/\operatorname{tr} G_j(\vartheta_j)$ and $B_n(\theta) = n^{-1}\sum_j G_j(\theta)/\operatorname{tr} G_j(\vartheta_j)$, assuming positive denominators. These two analysis matrices use the same historical weights. Define $\epsilon_{\mathrm{probe}} = \|C_n^{\mathrm{all}} - A_n\|_2$ and $\epsilon_{\mathrm{drift}} = \|B_n(\theta) - A_n\|_2$. Then

$$p^\top B_n(\theta) p \le \big(\|D_n\|_2 + \epsilon_{\mathrm{probe}} + \epsilon_{\mathrm{drift}}\big)\|p\|_2^2, \qquad U_n^\top p = 0. \tag{17}$$

Probe error includes self-normalization and conditioning on valid updates; drift error measures changing Jacobians. Both terms depend on the task distributions and boundary Jacobians.

A.3.2 A DIRECTION THAT NEVER ACCUMULATES

Let $d = 2$, $r = 1$, and consider the deterministic unit-trace sequence

$$S_1 = \operatorname{diag}(0.6, 0.4), \qquad S_j = \operatorname{diag}(0.4, 0.6) \quad (j \ge 2). \tag{18}$$

The first truncation keeps only e_1. Inductively, $\widetilde{C}_n = \operatorname{diag}(0.4 + 0.2/n, 0)$: at step $n \ge 2$, the candidate second eigenvalue is only $0.6/n$, strictly below $0.4 + 0.2/n$. In contrast, $C_n^{\mathrm{all}} = \operatorname{diag}(0.4 + 0.2/n, 0.6 - 0.2/n)$ has leading direction e_2 for every $n \ge 3$. Thus $\|D_n\|_2 = 0.6 - 0.2/n$ tends to 0.6, and BFP's worst residual sensitivity exceeds the full-history optimum by $0.2 - 0.4/n$ for $n \ge 3$. Repetition does not rescue a direction discarded before it can accumulate. This failure requires neither stochastic estimation nor model drift.

A.4 SCOPE AND RELATION TO PRIOR METHODS

Output-gradient protection (Farajtabar et al., 2020), activation subspaces (Saha et al., 2021), and adaptive constraint release (Deng et al., 2021) precede BFP. In vision–language learning, KeepLoRA uses pretrained and historical feature directions (Luo et al., 2026), NuSA-CL constrains updates through weight null spaces (Jo & Kim, 2026), and PCLR compresses adapter capacity (Meng et al., 2026). BFP fixes the model parameterization and reallocates a bounded functional-protection statistic. Its spectral optimum concerns that statistic in fixed parameter coordinates; the paired protocol evaluates adaptation and retention in the vision–language setting.

B EXPERIMENTAL DETAILS

B.1 BENCHMARKS AND TRAINING CONFIGURATION

The benchmarks measure task performance and scalability. The paired diagnostic in Appendices A.1, B.2, and B.3 tests PIPL by varying protection at a shared checkpoint.

B.1.1 BENCHMARKS, SPLITS, AND EVALUATION VOCABULARIES

VLCL retrieval and pretrained retention. The standard stream follows Flickr30K, MS-COCO, Oxford-IIIT Pets, Lexica, Simpsons, WikiArt, Kream, and Sketch (Liu et al., 2025). We use the benchmark's captions and released split identifiers: Flickr30K's 1K and COCO's 5K test galleries; the released Pets and Lexica tests; 80/20 training/test partitions for Simpsons, WikiArt, and Sketch; and the 50/50 Kream partition. Original captions are preserved and grouped by image, and duplicate-image overlaps are audited before training. Any split reconstructed from a fixed seed is identified in the data release and labelled as a reconstructed-split result.

We report $R@1$ for each retrieval direction separately. Recall uses the entire fixed gallery, with any ground-truth caption counted as correct for an image query. Averaging over domains gives equal weight to each domain. Evaluation-only classification uses ImageNet-1K, CIFAR-100, Flowers-102, DTD, Food-101, and Stanford Cars. The single prompt and test sets are fixed in Appendix B.3.1. A reproduction using a source method's prompt ensemble is a separately labelled result; the common-prompt rerun uses the same prompts for every method.

MTIL and class-incremental recognition. MTIL uses the standard order Aircraft, Caltech101, CIFAR100, DTD, EuroSAT, Flowers102, Food101, MNIST, OxfordPet, StanfordCars, and SUN397 (Zheng et al., 2023; Luo et al., 2026). We retain the benchmark-specific training/test manifests and class-name templates. Full-shot training uses the full available training split; five-shot training samples five images per class once per run, using identical selections across methods (Jo & Kim, 2026). Neither test images nor future-domain training images enter current-task optimization or probing. MTIL evaluation supplies the candidate vocabulary of the dataset being evaluated. This restricted vocabulary provides dataset identity during evaluation, while BFP uses a single model across datasets.

For a task-agnostic check, X-TAIL uses the ten MTIL domains other than CIFAR-100, with the union of seen and unseen class vocabularies at inference (Zhang et al., 2026). We retain the released class-ID mapping and evaluate each domain against that common candidate set. Transfer, Average, and Last use the same dataset-wise aggregation below with $T = 10$, but all entries come from union-vocabulary evaluation. These results are kept separate from restricted-vocabulary MTIL; no domain label selects a classifier or expert. The comparison set is frozen CLIP, sequential LoRA, ZSCL, MoE-Adapters, and Pi-CCA.

For longer sequences of new classes, CIFAR-100 is divided into 10, 20, or 50 equal increments, following the ZSCL/NuSA-CL protocol (Zheng et al., 2023; Jo & Kim, 2026). A shared class permutation fixes the 10-, 5-, or 2-class stages. Evaluation uses the union of all classes encountered so far, without the test image's task identity. The exact released class order is used for the canonical comparison, and one shared fixed permutation is used for the common-budget comparison. Its average incremental accuracy is a different quantity from MTIL's Average.

On classification tasks, BFP replaces the pairwise contrastive task loss by cross-entropy over image–text class logits, while retaining the same adapters, functional probes, and projected optimizer. During MTIL training the softmax contains current-domain classes; during class-incremental training it contains all seen classes. Probe pairs consist of a current training image and its ground-truth class prompt. This classification extension replaces the contrastive objective in Eq. (2) with a class-level softmax.

Boundary-density stress test and diagnostic split. The 32-stage VLCL variant partitions each domain's training images into four disjoint groups, retaining all captions of each image in its group. The eight-stage stream uses 1,000 updates per domain, and the 32-stage stream uses 250 per group; both total 8,000 updates. We compare contiguous domain groups with a seeded interleaving that

preserves each domain's internal group order. The variants match sample availability and optimizer-update counts, with minibatch order determined by the partition. Test galleries stay fixed at the original eight domains. This construction varies boundary density within those domains. Additional boundaries add probing and SVD work to the matched update budget.

The primary paired diagnostic is a separate seven-domain parent experiment: Sketch is never trained by either parent. It uses the global image-deduplicated 80/20 split and seeds $0, \ldots, 9$ in Appendix B.2; the standard benchmark retains the published VLCL galleries. Checkpoints at $t = 1, 4, 7$ share the same 200-step Sketch adaptation schedule within a run. Diagnostic labels, test scores, and branch states never feed back into the parent trajectory.

B.1.2 Metrics and uncertainty

Let $a_{i,j}$ denote top-1 accuracy on MTIL dataset j after training stage i, with $T = 11$. Following the dataset-wise aggregation used by KeepLoRA (Luo et al., 2026),

$$\text{Transfer} = \frac{1}{T-1} \sum_{j=2}^{T} \frac{1}{j-1} \sum_{i=1}^{j-1} a_{i,j}, \tag{19}$$

$$\text{Average} = \frac{1}{T^2} \sum_{i=1}^{T} \sum_{j=1}^{T} a_{i,j}, \qquad \text{Last} = \frac{1}{T} \sum_{j=1}^{T} a_{T,j}. \tag{20}$$

The frozen model's $a_{0,j}$ is excluded from these sums; its aggregate scores appear as a reference row. Transfer first averages a dataset's evaluations before it is learned, then averages datasets; pooling all upper-triangular matrix entries would weight later datasets more heavily. Average includes both seen and unseen datasets, whereas Last averages all datasets after the final stage.

For each evaluation-only classification dataset d, pretrained performance drop is $\text{PD}_d = a_{0,d} - a_{T,d}$ in percentage points. The source-protocol columns report ImageNet-1K and CIFAR-100 drops; the common-budget PD_6 is the unweighted mean over the six evaluation-only datasets. These datasets serve as evaluation-only tests in VLCL; MTIL includes some of them as training domains. For CIFAR-100 class increments, we report final all-seen accuracy and mean all-seen accuracy over stages, separately from MTIL Average.

PIPL uses the normalized trapezoidal AUC difference Δ_t, historical retention change, and pretrained retention change already defined in Appendix B.3. Primary decisions retain its seven simultaneous one-sided bounds, ten independent paired runs, and fixed 0.01 margins. Our fixed-order benchmark evaluations also use seeds $0, \ldots, 9$ and report means over complete runs; Table 8 additionally reports sample standard deviations. Backbone and boundary-density comparisons are descriptive analyses. Full runs are the independent replications; standard deviations summarize variation across those runs.

B.1.3 Comparator selection and fairness

Literature-based comparison sets. The retrieval set follows the methods compared in C-CLIP's Table 3 and Pi-CCA's Table 2: EWC (Kirkpatrick et al., 2017), ZSCL (Zheng et al., 2023), Mod-X (Ni et al., 2023), DKR (Cui et al., 2024), MoE-Adapters (Yu et al., 2024), C-CLIP (Liu et al., 2025), and Pi-CCA (Zhang et al., 2026). MTIL retains ZSCL, MoE-Adapters, and Pi-CCA, and adds the directly relevant low-rank methods InfLoRA (Liang & Li, 2024), SD-LoRA (Wu et al., 2025), KeepLoRA (Luo et al., 2026), and NuSA-CL (Jo & Kim, 2026), informed by KeepLoRA's Table 2 and NuSA-CL's Table 1. The CIFAR-100 extension compares sequential LoRA, EWC, ZSCL, InfLoRA, SD-LoRA, KeepLoRA, and NuSA-CL under the same class schedule. Frozen CLIP and sequential LoRA (Radford et al., 2021; Hu et al., 2022) are included throughout. Full fine-tuning provides a reference with a larger trainable parameter count. InfLoRA and SD-LoRA use the CLIP adaptations evaluated by KeepLoRA/NuSA-CL. Moving an MTIL method to CIFAR-100 is likewise labelled a benchmark adaptation.

Mechanism comparisons in fixed coordinates. The mechanism comparisons use BFP's two rank-eight projection adapters, AdamW convention, training minibatches, and probe schedule. The five parent variants are no protection, growing protection, a basis frozen after the first task, a fixed random orthonormal basis, and BFP. The last three share the rank budget $r = 16$. Growing

protection is an internal ablation that removes BFP's rank cap. These comparisons use only the seven-domain ViT-B/16 diagnostic stream. Table 11 defines their late-checkpoint comparison; norm replay, orientation/deletion, and optimizer-reset controls at $t = 7$ appear in Table 13.

Architecture-preserving comparisons. Each external method retains the layers it adapts, its teacher or routing mechanism, and its merge/reinitialization rules. KeepLoRA and NuSA-CL reparameterize and merge low-rank factors between tasks; BFP retains fixed factor coordinates. Checkpoint tags and weight hashes identify the pretrained weights, including Pi-CCA's OpenCLIP backbone. Reproductions using a different checkpoint remain separate from the common-OpenAI-checkpoint reruns.

Canonical reproductions preserve each source's data, prompts, schedule, and native hyperparameters, and are reported separately. Harmonized comparisons then fix the checkpoint, split, task order, evaluation vocabulary, prompt policy, and total update budget. They preserve each method's architecture and method-specific losses. ZSCL retains external-reference distillation and its teacher, while MoE-Adapters retains its experts and routing. Parameter counts and wall-clock costs vary with these method-specific components.

Protection-state measurements are restricted to BFP and growing protection on the seven-domain ViT-B/16 stream (Table 14). At $t = 1, 4, 7$, record realized rank, resident protection-state bytes, mean projection time per update during that domain, and boundary-refresh time including probing and SVD. Use the same GPU and precision, synchronized timings, and no feature cache for this measurement. Diagnostic branches and test evaluation are excluded from these component timings. The measurements quantify the two parents' protection overhead.

Related evaluation settings. Our comparisons focus on Euclidean CLIP models and shared evaluation vocabularies. Other recent settings include Lorentz-space MERU/HyCoCLIP models in HMCL (Liu et al., 2026) and dual-forgetting class-incremental learning in DFA-MoE (Kang et al., 2026). Generative benchmarks such as CoIN (Chen et al., 2024), with methods including PCLR (Meng et al., 2026) and SEFE (Chen et al., 2025), use generative output probes and task-specific scorers.

B.1.4 Backbones, budgets, and implementation choices

Backbone and parameter budget. The primary model is OpenAI CLIP ViT-B/16, the common backbone in the closest projection/low-rank comparisons (Luo et al., 2026; Jo & Kim, 2026). Robustness runs use the OpenAI ViT-B/32 and ViT-L/14 checkpoints, all at 224×224 input resolution. Each model retains its own published tokenizer, image normalization, embedding dimension, and pretrained temperature. BFP adds two rank-eight residual projections with scale one and no taskwise merging, reinitialization, or expansion. Its trainable dimension is

$$d = 8(p_{\mathrm{img}} + p_{\mathrm{txt}} + 2q), \tag{21}$$

giving 18,432 parameters for either base model and 26,624 for ViT-L/14. The protection rank remains 16 across backbones. All baseline comparisons are within a backbone; pretrained scores are recorded before any updates.

Fixed training configuration. The fixed retrieval/diagnostic configuration is Table 4: AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, learning rate 10^{-4}, weight decay 0.01, batch size 256, and 1,000 updates per retrieval domain, with no clipping or stochastic augmentation. Model computations use FP32, and sketch accumulation, SVD, and projection use FP64. The full optimizer proposal, including decay, is projected once; moments are updated from raw gradients and persist through task boundaries. No dense $d \times d$ projector is formed. Because only the final projections train, deterministic frozen-backbone features may be cached for this controlled track, with identical caching for every compatible control. Protection-overhead measurements use the uncached path.

The common budget for MTIL is 1,000 updates per full-shot domain or 500 per five-shot domain, with batch size 16, following NuSA-CL's iteration budgets (Jo & Kim, 2026). The CIFAR-100 extension uses batch size 128 and a fixed total 10,000 updates, allocated equally across 10/20/50 stages; this update-matched extension is labelled separately from canonical source-schedule reproductions. BFP keeps learning rate 10^{-4}, $r = 16$, and $m = 64$ as its fixed configuration in these extensions. These

extensions retain the same protection recurrence. Probe sampling uses only the current training split and releases raw examples after sketch construction.

All reported BFP configurations fix adapter rank eight, protection rank $r = 16$, and probe budget $m = 64$. The diagnostic protocol and retention margins are fixed independently of test scores.

Hardware and reproducibility. The hardware configuration specifies one NVIDIA A100 80GB GPU per run and at least 128GB host RAM. Effective batch size is preserved if memory requires microbatching. Contrastive microbatch accumulation requires the full logical batch's negatives, for example through differentiable gathered embeddings. The internal protection-overhead comparison uses one common device type. The reproducibility record includes GPU count/model, precision, PyTorch/CUDA/driver versions, source commits, checkpoint hashes, split and order manifests, prompts, seeds, and failed runs. Paired run-level scores and actual protection ranks are retained to check aggregation, confidence bounds, and the recurrence.

B.2 Deterministic intervention protocol

B.2.1 Data, parents, and checkpoints

The diagnostic construction uses the eight caption domains of VLCL (Liu et al., 2025). Sketch is reserved for diagnostic adaptation; the parent domains are Flickr30K, COCO, Pets, Lexica, Simpsons, WikiArt, and Kream. This diagnostic uses a seven-domain parent stream with Sketch held out; the standard VLCL benchmark uses all eight domains. Group records globally by the SHA-256 digest of the image file, keeping all associated captions together. Remove from parent domains any image digest present in Sketch. Across the resulting eight-domain union, rank the N unique image groups by SHA-256 of the string `26042:split:digest`. Assign the first $\lfloor 0.8N \rfloor$ groups to training and the rest to evaluation, then restrict this assignment to each domain. This global assignment keeps each image in one split across domain labels. The split manifest records image digests, domain names, captions, and assignments.

Ten independent run seeds $b \in \{0, \dots, 9\}$ determine adapter initialization, a permutation of the seven parent domains, training batches, and boundary probes. Separate random streams keep later minibatches independent of skipped probes. Within each run, BFP and growing protection share their random schedules. Training samples distinct images within each batch and one of each image's captions uniformly; shuffle image groups on each pass and drop incomplete batches. Each parent executes $1{,}000$ updates per domain with Table 4.

Save checkpoints immediately after the protection refresh at boundaries $t = 1, 4, 7$. Each contains parameters, model buffers, optimizer moments and counter, the ordered protection basis, and random states. The diagnostic task uses $K = 200$ updates and evaluation steps $\{0, 20, \dots, 200\}$. Its data schedule is generated once per run and reused at all checkpoints and across both parents. All branches within a checkpoint start from the identical saved state. Their bases remain fixed, and no diagnostic probes refresh memory. The parent resumes from its saved state after the branch measurements.

B.2.2 Release and reference-norm replay

For a checkpoint with k protected directions ordered by stored eigenvalue, the full basis is U and the half-release basis is $U^- = U_{:,1:\lfloor k/2 \rfloor}$. This single release rule is fixed at $t = 1, 4, 7$. With $k = 0$, the two primary branches coincide and the release effect is zero. The additional norm, orientation, and optimizer-reset controls below are restricted to the late checkpoint $t = 7$.

First execute the full-protection reference branch and record the actual displacement norm $b_s = \|\theta_s^{\text{full}} - \theta_{s-1}^{\text{full}}\|_2$ at each diagnostic step. In a separate half-release control, compute the complete AdamW proposal u_s^{ctl} from that branch's own state and put $v_s = P_{U^-} u_s^{\text{ctl}}$. Its displacement is

$$d_s^{\text{ctl}} = \begin{cases} -b_s v_s / \|v_s\|_2, & b_s > 0,\ \|v_s\|_2 > 0, \\ 0, & b_s = 0. \end{cases} \qquad \theta_s^{\text{ctl}} = \theta_{s-1}^{\text{ctl}} + d_s^{\text{ctl}}. \tag{22}$$

For $b_s > 0$ and $v_s = 0$, the matched control is infeasible: terminate the control and record the failure step. The primary unrescaled branches remain valid. Optimizer moments are updated even when $b_s = 0$; no extra weight decay follows (22). Record the realized norm mismatch after parameter

rounding. The control replays the reference trajectory's step lengths while its update directions evolve from its own model state.

B.2.3 Orientation and state controls

At the late checkpoint $t = 7$, two additional controls retain $k^- = \lfloor k/2 \rfloor$ constraints: a Haar-random k^--dimensional subspace of $\mathbb{R}^d$, and a uniformly sampled k^--column subset of U. Generate the former by thin QR of an independent Gaussian matrix, with positive diagonal entries in its triangular factor. Freeze each basis for all K steps. Compare both with the norm-matched half-release branch: the first changes arbitrary orientation at fixed rank, and the second changes which historical directions are retained. All three use the same full-protection reference norms and infeasibility rule.

At $t = 7$, repeat the full/half checkpoint pair with both optimizers reset to fresh moments and step counter zero, keeping the learned parameters and all other states fixed. This secondary comparison tests optimizer-state dependence. Differences between checkpoints reflect the joint evolution of the model and constraints, evaluated on the same diagnostic task.

B.2.4 Fixed-coordinate parent controls

All five parents share initialization, minibatches, probe draws, and the same two adapters. They train the first domain without protection. No protection continues with an empty basis; growing protection retains all numerically nonzero directions. The frozen-basis control retains the leading at most 16 directions after the first domain. The random-basis control draws an independent Haar basis of that same realized rank after the first domain. Both bases then remain fixed. BFP refreshes its basis with budget 16. Probe collection follows the common schedule even when a control does not use its sketch.

B.2.5 Protection-overhead measurements

State size counts only resident (U, Λ, n) tensors after refresh, in MiB (2^{20} bytes). Projection time averages synchronized applications of $u - U(U^\top u)$ during domain t, using U_{t-1}; refresh time includes probing, normalization, and SVD producing U_t. Both use the same A100, precision, and uncached path. These are component costs, excluding diagnostic branches and test evaluation.

B.3 Metrics, uncertainty, and decision rules

B.3.1 Fixed scores and aggregation

For an evaluation gallery $\mathcal{G}$, define $Q(\theta; \mathcal{G})$ as the arithmetic mean of image-to-text and text-to-image Recall@1. An image query succeeds if its highest-scoring caption belongs to that image; a caption query succeeds if its highest-scoring image is its source. Use all evaluation images and their captions as candidates, with cosine similarity and lexicographic identifier tie breaking. Average within each query direction before averaging the two directions. Then $S_t^V(s) = Q(\theta_{t,s}^V; \mathcal{G}_\star)$ and

$$\widehat{\mathrm{AUC}}_t(V) = \frac{1}{K} \sum_{j=1}^{10} 20 \, \frac{S_t^V(20(j-1)) + S_t^V(20j)}{2}, \qquad K = 200. \tag{23}$$

The old-task score R_t^V averages $Q(\theta_{t,K}^V; \mathcal{G}_i)$ equally over the t previously learned domains. Repeated examples never increase a domain's weight.

The pretrained score Z_t^V is the unweighted mean of sample top-1 accuracies on ImageNet-1K validation, CIFAR-100 test, Oxford Flowers-102 test, DTD test split 1, Food-101 test, and Stanford Cars test. These evaluation datasets follow the classification suite of VLCL (Liu et al., 2025); none provides training examples to either parent. For every class, encode the single prompt `a photo of a {class name}.` with the current text branch, replace underscores in class names by spaces, normalize the resulting embedding, and select the class of highest image–text cosine similarity. Use the released class ordering to break ties. This fixes the classifier without prompt search. Historical and pretrained evaluations occur at diagnostic steps 0 and K and never change branch updates.

B.3.2 PRIMARY FAMILY AND SIMULTANEOUS LOWER BOUNDS

Write $\Delta^G_{b,t}$ and $\Delta^B_{b,t}$ for the half-release effects in run b of growing protection and BFP. At the late checkpoint $t = 7$, define $dR^a_b = R^{a,-}_{b,7} - R^a_{b,7}$ and $dZ^a_b = Z^{a,-}_{b,7} - Z^a_{b,7}$ for parent $a \in \{G, B\}$. The seven primary run-level quantities are

$$Y_b = \big(\Delta^G_{b,7}, dR^G_b, dZ^G_b,\ \Delta^B_{b,7}, dR^B_b, dZ^B_b,\ \Delta^G_{b,7} - \Delta^G_{b,1}\big). \tag{24}$$

The unit of replication is an independent full run. For component j, let $\bar{Y}_j$ and s_j be its mean and sample standard deviation over the $N = 10$ paired runs. Use the one-sided lower bound

$$L_j = \bar{Y}_j - t_{N-1,\,1-0.05/7}\frac{s_j}{\sqrt{N}}, \qquad s_j^2 = \frac{1}{N-1}\sum_{b=0}^{N-1}(Y_{b,j} - \bar{Y}_j)^2. \tag{25}$$

Here $t_{\nu,p}$ is the p-quantile of Student's t distribution with ν degrees of freedom. Bonferroni adjustment covers the seven prespecified components. These nominal parametric intervals assume approximately Gaussian, independent run-level summaries.

Growing protection is classified as excessive only if $L_1 > 0.01$, $L_2 \geq -0.01$, and $L_3 \geq -0.01$; BFP uses the same rule on L_4, L_5, L_6. A history-associated increase in the growing parent's release effect requires $L_7 > 0.01$. All three margins are absolute score differences: one percentage point on the $[0, 1]$ scale. The protocol fixes these margins in advance. Classification as excessive protection requires all three corresponding bounds to pass; the early-to-late contrast measures the change in release effect.

The middle checkpoint, fixed-coordinate parent comparisons, and late-checkpoint controls are descriptive secondary analyses (Appendix C.4); the seven-component primary family remains unchanged. Matched-norm comparisons require feasible controls in all ten runs. Primary decisions likewise require ten complete paired runs, with any failed run reported and retained in the run count.

B.4 PRELIMINARY-STUDY PROTOCOL

B.4.1 PASSIVE NEXT-DOMAIN PROBING

The study uses CLIP ViT-B/16 with the two rank-eight projection adapters and joint $d = 18{,}432$-dimensional parameter vector specified in Appendix A.1.1. The growing-protection parent retains every realized nonzero historical output-sensitivity direction under the numerical-rank rule in Appendix A.1.2. Its seven parent domains are Flickr30K, COCO, Pets, Lexica, Simpsons, WikiArt, and Kream; the shared data construction and run schedules are given in Appendix B.2.1.

Let $s \in \{0, \ldots, 9\}$ index runs. After training domain t and refreshing the protection basis, we froze the checkpoint $\theta_{s,t}$ and basis $U_{s,t} \in \mathbb{R}^{d \times k_{s,t}}$. At each $t \in \{1, \ldots, 6\}$, the probes use the next scheduled domain in that run:

$$g_{s,t,b} = \nabla_\theta \ell(\theta_{s,t}; \mathcal{B}_{s,t+1,b}), \qquad b = 1, \ldots, 64, \tag{26}$$

where ℓ is the contrastive loss in Eq. (2). During these measurements, the model, optimizer, protection basis, and parent training schedule remain fixed. Each of the 60 run–boundary pairs contributes 64 probes, giving 640 per history length and 3,840 overall. The index b counts minibatches. The paired intervention in Section 4 evaluates adaptation on the reserved Sketch domain at boundaries 1, 4, and 7.

B.4.2 INTERCEPTION, RANK REFERENCE, AND AGGREGATION

Equation (1) uses the raw task gradient at the frozen checkpoint and suppresses the run index. For nonzero gradients, $\rho_{s,t,b} \in [0, 1]$ is the squared norm captured by the orthogonal projector $U_{s,t}U_{s,t}^\top$, divided by total gradient energy. For a Haar-random rank-$k_{s,t}$ basis $\widetilde{U}$ and fixed nonzero g, rotational symmetry gives

$$\mathbb{E}[\widetilde{U}\widetilde{U}^\top] = \frac{k_{s,t}}{d} I_d, \qquad \mathbb{E}\left[\frac{\|\widetilde{U}^\top g\|_2^2}{\|g\|_2^2}\right] = \frac{k_{s,t}}{d}. \tag{27}$$

The first identity follows because the expected projector is isotropic and has trace $k_{s,t}$. We subtract the reference for each checkpoint before pooling:

$$e_{s,t,b} = \rho_{s,t,b} - \frac{k_{s,t}}{d}, \qquad \bar{r}_t = \frac{1}{10}\sum_{s=0}^{9} \frac{k_{s,t}}{d}. \tag{28}$$

The column-top values in Figure 2(a) are $\bar{r}_t$. Equal probe counts give every run equal weight. At each t, medians and quantiles are computed over the 640 individual probe values; the median excess is the median of $e_{s,t,b}$. The amplification reported in Appendix C.1 is the pooled median of $\rho_{s,t,b}$ divided by $\bar{r}_t$.

B.4.3 Semantic-affinity measurement

Using frozen pretrained CLIP features, let $\bar{z}_0(\mathcal{B})$ denote the normalized mean image–text representation of a minibatch and $\mu_{s,j}$ the corresponding prototype of historical domain j in run s. The recorded affinity is

$$a_{s,t,b} = \max_{j \le t} \cos\big(\bar{z}_0(\mathcal{B}_{s,t+1,b}), \mu_{s,j}\big). \tag{29}$$

Affinity quartiles are formed separately at each history length by pooling its 640 probes, yielding 160 probes per quartile. We then compute median excess within each group.

B.4.4 Construction of Figure 2

Quantile surface. For each history length, panel (a) evaluates empirical quantiles at $0.02, 0.04, \ldots, 0.98$ using linear interpolation between sorted probe values. Cells are centered at these quantiles and at the discrete history lengths. The white horizontal line and circles mark the median. The purple color scale uses a power normalization with exponent 0.82; its colorbar reports interception ratios on their original scale.

Affinity map. Panel (b) groups the recorded affinity and vertically jittered history positions into hexagons with grid size $(33, 13)$. Each displayed hexagon contains at least four probes and is colored by their median excess. The stored jitter lies within 0.5 of the integer history length and is used only for this display. Diamonds lie at each history length's median affinity; the accompanying labels show its pooled median excess in percentage points, rounded to one decimal place.

Density contours and numerical precision. Gray contours use a 54×49 histogram of affinity and display position, smoothed by a Gaussian filter with standard deviation 1.1 bins. Contour levels are the 45th, 68th, and 84th percentiles of positive smoothed cell counts. Numerical summaries use the original discrete history lengths and unrounded probe values. Percentages and percentage point differences are rounded only for reporting.

B.5 Source protocols and reporting conventions

Source-reported baseline entries retain the published precision; BFP, Pi-CCA, common-budget entries, and the five-shot extensions below are our evaluations. Benchmark tables report run means, with sample SD in Table 8. Dashes denote unavailable results. Shading identifies BFP. In Tables 1–2, boldface and underlining mark the best and second-best adapting methods per metric within each protocol; Frozen CLIP is a fixed reference excluded from ranking. Green subscripts in Table 2 show BFP's absolute gains in percentage points over the strongest baseline in the same column and panel. Appendix benchmark tables use boldface for the best adapting method.

VLCL retrieval and retention (Table 1). I2T/T2I report final eight-domain macro $R@1$, giving equal weight to each domain. I1K/C100 denote ImageNet-1K/CIFAR-100; PD_6 averages the six held-out classification datasets. These performance drops are in percentage points. Source-reported baseline entries reproduce C-CLIP Tables 3–4 (Liu et al., 2025), including its "MOE-CL" entries under MoE-Adapters. BFP uses the corresponding 40-epoch, batch-1,024 schedule. Pi-CCA entries in the source-protocol columns are our evaluations under this protocol. The common budget uses 1,000 updates/domain and batch 256. Source I1K/C100 drops and common-budget PD_6 remain separate metrics.

MTIL classification (Table 2). Transfer uses unseen domains, Average includes all stages and domains, and Last evaluates the final model; Appendix B.1.2 gives the aggregation formulas. Full fine-tuning denotes continual full-model fine-tuning. Panel (a) follows NuSA-CL with 1,000/500 updates per domain for full-/five-shot training and batch 16. Panel (b) follows the KeepLoRA pipeline with native full-shot schedules and a matched five-shot extension; KeepLoRA's native schedule uses batch 64 and at most ten epochs/domain. Full-shot baselines are from NuSA-CL Table 1 and KeepLoRA Table 2; five-shot LoRA/InfLoRA/NuSA-CL in panel (a) are from NuSA-CL Table 2 (Jo & Kim, 2026; Luo et al., 2026). The five-shot scores in panel (b) and BFP are our evaluations. BFP in panel (a) reports the same protocol-matched runs as Table 8, rounded to one decimal place. BFP retains two rank-eight projection adapters, $r = 16$, and $m = 64$. The two InfLoRA implementations remain separate. ZSCL retains its reference-data/teacher resources and MoE-Adapters its routing; parameter counts and auxiliary-resource budgets therefore vary across methods.

Paired diagnostic (Table 3). Growing protection is our uncapped BFP control (Appendix A.1.2). The separate seven-domain parent experiment uses ten paired runs and $K = 200$ diagnostic updates on held-out Sketch. Panel (a) reports mean $\pm$ sample SD; boldface and underlining identify the higher and lower means, respectively, while SDs remain unranked. Superscript $-$ denotes half release. Panel (b) reports unranked release effects and simultaneous one-sided lower bounds L, with $dR_7 = R_7^- - R_7$ and $dZ_7 = Z_7^- - Z_7$. The late decision requires $L(\Delta_7) > 1$, $L(dR_7) \geq -1$, and $L(dZ_7) \geq -1$ pp; the history contrast requires $L(\Delta_7 - \Delta_1) > 1$ pp. Growing protection has $\Delta_7 - \Delta_1 = 5.50$ pp and $L(\Delta_7 - \Delta_1) = 4.33$ pp. All seven bounds share the correction in Appendix B.3; earlier effects are descriptive.

Common-budget Pi-CCA (Zhang et al., 2026) entries use the shared OpenAI checkpoint and evaluation protocol, separately from the original paper's OpenCLIP scores. All backbone comparisons use eight domains and matched update budgets.

B.6 Direction-level protection-utility protocol

Paired diagnostic. At $t = 7$, let $U = [u_1, \ldots, u_k]$ be a parent's protection basis and let $U^{(-i)}$ remove column u_i, preserving the remaining column order. Within each parent and run, the full-basis and single-direction-release branches start from identical model parameters, optimizer state, random state, and minibatch schedule. Each branch receives $K = 200$ held-out Sketch updates under the setup of Table 3. The analysis covers ten paired runs and all 4,431 growing-protection and 160 BFP directions.

Direction-level measurements. Using the adaptation AUC in Eq. (3), the gain from releasing u_i is

$$G_i = \mathrm{AUC}_7\Big(U^{(-i)}\Big) - \mathrm{AUC}_7(U). \tag{30}$$

Retention damage is the largest nonnegative drop in historical retrieval R_7 or pretrained zero-shot accuracy Z_7, evaluated after the diagnostic updates:

$$D_i = \max\Big\{0,\ R_7(U) - R_7\Big(U^{(-i)}\Big),\ Z_7(U) - Z_7\Big(U^{(-i)}\Big)\Big\}. \tag{31}$$

Both quantities are in percentage points. Each direction also carries its stored protection eigenvalue λ_i and future-gradient overlap score $q_i \in [0, 1]$.

Utility concentration. For each parent, we pool direction records across runs and define the utility region as $\mathcal{U} = \{i : G_i > 0.8,\ D_i < 0.15\}$. Its share of summed positive single-direction gains is

$$S_{\mathcal{U}} = \frac{\sum_{i \in \mathcal{U}} \max(G_i, 0)}{\sum_i \max(G_i, 0)}. \tag{32}$$

Visual encoding. Figure 5 displays every direction with shared axes, color normalization, and bubble-area scaling. Bubble area is an affine function of $\log(1 + \lambda_i)$ with a positive minimum area. Gaussian kernel density estimates use bandwidth factors 0.15 and 0.30 for growing protection and BFP, respectively; contours mark the 70th, 84th, and 93rd percentiles of positive density values on the evaluation grid.

C Additional Experimental Results

C.1 Preliminary-study numerical summaries

The measurements and affinity groups follow Appendix B.4.

Table 5: **Interception summaries at all six history lengths.** Each row pools 640 probes from ten runs. Q_p is the empirical pth percentile; pp denotes percentage points.

t	Rank range	$\bar{r}_t$ (%)	Median ρ (%)	Q_{10}–Q_{90} (%)	Median e (pp)	Amplification
1	61–64	0.34	8.55	5.40–13.10	8.21	25.3×
2	123–128	0.68	11.60	7.30–19.24	10.91	17.1×
3	185–191	1.02	15.45	9.25–25.16	14.42	15.2×
4	249–255	1.37	23.64	13.16–37.97	22.28	17.3×
5	312–320	1.71	33.71	18.59–48.25	32.01	19.8×
6	377–384	2.06	39.90	25.32–57.94	37.85	19.3×

Table 5 gives the complete history-level summaries. Median interception is 15.2–25.3 times the mean rank-only expectation. From $t = 1$ to $t = 6$, the 90th percentile rises from 13.10% to 57.94%, and the 10th–90th percentile spread widens from 7.70 to 32.62 percentage points. At $t = 6$, median excess across the four affinity quartiles is 25.14, 33.89, 40.20, and 51.73 percentage points, respectively. The quartile medians increase with affinity at each of the six history lengths.

C.2 Benchmark comparisons

C.2.1 Domain-level retrieval

Table 6: **Final retrieval on the original eight-domain VLCL protocol.** Final R@1 (%, ↑); baselines are from C-CLIP Table 3 (Liu et al., 2025), and BFP is evaluated under the same protocol. The macro mean weights domains equally. These results correspond to the source-protocol columns of Table 1.

Method	Flickr	COCO	Pets	Lexica	Simp.	WikiArt	Kream	Sketch	Mean
Image-to-text R@1 (↑)									
Frozen CLIP	35.80	10.40	1.96	3.99	2.62	4.40	9.41	1.70	8.79
EWC	60.38	31.58	12.23	20.06	15.13	37.40	33.63	7.93	27.29
ZSCL	66.79	37.15	13.34	24.99	20.98	40.67	39.23	6.88	31.25
Mod-X	61.62	33.94	11.38	18.04	18.18	39.39	35.94	6.95	28.18
MoE-Adapters	63.63	34.11	12.45	23.03	17.03	41.09	38.67	7.27	29.66
DKR	63.60	28.98	11.04	21.74	19.05	38.02	37.92	7.06	28.43
C-CLIP	84.40	56.92	19.73	42.65	25.43	45.89	42.07	9.55	40.83
BFP (ours)	**85.90**	**58.20**	**21.10**	**44.10**	**27.00**	**47.60**	**44.10**	**10.88**	**42.36**
Text-to-image R@1 (↑)									
Frozen CLIP	55.88	28.83	10.28	19.59	13.43	16.34	18.30	4.25	20.86
EWC	60.30	31.16	13.91	20.14	15.52	28.34	30.34	12.23	26.49
ZSCL	65.52	39.85	14.22	25.37	19.91	34.76	34.99	11.98	30.83
Mod-X	61.16	35.37	13.01	21.84	17.24	30.98	28.62	10.66	27.36
MoE-Adapters	63.17	40.31	15.45	24.09	16.36	35.78	34.04	11.02	30.03
DKR	62.72	37.73	14.94	23.60	20.00	31.31	30.91	10.26	28.93
C-CLIP	73.74	42.82	17.91	41.47	24.32	45.27	43.57	14.67	37.97
BFP (ours)	**75.10**	**44.60**	**19.20**	**43.10**	**25.90**	**46.90**	**45.10**	**14.90**	**39.35**

C.2.2 CLASS-INCREMENTAL RECOGNITION

Table 7: **CIFAR-100 class-incremental recognition.** ViT-B/16; accuracy in percent (↑). Each stage group reports final all-seen accuracy (Last) and average incremental accuracy (Avg.). Panel (a) uses native schedules, with baseline scores from NuSA-CL Table 3 (Jo & Kim, 2026) and ZSCL Table 6 (Zheng et al., 2023). Panel (b) fixes the total update budget.

	10 stages		20 stages		50 stages	
Method	Last	Avg.	Last	Avg.	Last	Avg.
(a) Native training schedules						
Frozen CLIP (Radford et al., 2021)	65.92	74.47	65.74	75.20	65.94	75.67
Continual full fine-tuning	53.23	65.46	43.13	59.69	18.89	39.23
ZSCL (Zheng et al., 2023)	73.65	82.15	69.58	80.39	67.36	79.92
NuSA-CL (Jo & Kim, 2026)	74.51	80.25	73.84	81.03	71.85	80.19
BFP (ours)	**75.72**	**82.98**	**75.01**	**82.28**	**73.06**	**81.42**
(b) Common budget: 10,000 updates total, batch size 128						
Frozen CLIP (Radford et al., 2021)	65.88	74.55	65.88	75.11	65.88	75.64
Sequential LoRA (Hu et al., 2022)	67.42	76.31	65.93	75.08	62.86	73.02
EWC (Kirkpatrick et al., 2017)	68.71	77.46	67.12	76.25	64.40	74.16
ZSCL (Zheng et al., 2023)	73.31	82.02	70.86	80.31	67.52	78.47
InfLoRA (Liang & Li, 2024)	72.96	81.07	71.38	80.16	68.93	78.82
SD-LoRA (Wu et al., 2025)	72.48	80.71	70.62	79.41	67.88	77.64
KeepLoRA (Luo et al., 2026)	74.34	82.31	72.81	81.22	70.36	80.01
NuSA-CL (Jo & Kim, 2026)	74.82	81.72	73.56	80.94	71.43	79.86
BFP (ours)	**76.03**	**83.17**	**74.96**	**82.39**	**72.94**	**81.51**

Native source scores and common-budget evaluations are kept separate. The source Frozen CLIP Last values are retained as printed; the common-budget panel uses one fixed class permutation. InfLoRA, SD-LoRA, KeepLoRA, and NuSA-CL are CLIP benchmark adaptations.

C.2.3 COMMON-BUDGET CLASSIFICATION

Table 8: **Classification under the common OpenAI ViT-B/16 protocol.** Full-shot/five-shot MTIL use 1,000/500 updates per domain with batch size 16; X-TAIL uses the full-shot budget on ten domains. All values are mean ± sample SD over ten runs (%, ↑). Tr./Avg./Last follow Appendix B.1.2. MTIL uses dataset-specific vocabularies; X-TAIL uses the union of seen and unseen vocabularies.

	MTIL full-shot			MTIL five-shot			X-TAIL		
Method	Tr.	Avg.	Last	Tr.	Avg.	Last	Tr.	Avg.	Last
Frozen CLIP	69.39 ±0.00	65.29 ±0.00	65.29 ±0.00	69.39 ±0.00	65.29 ±0.00	65.29 ±0.00	44.20 ±0.00	41.64 ±0.00	41.64 ±0.00
Sequential LoRA	63.50 ±0.61	69.60 ±0.53	79.00 ±0.64	60.20 ±0.72	66.10 ±0.61	71.70 ±0.69	41.20 ±0.63	45.60 ±0.58	49.30 ±0.66
ZSCL	67.90 ±0.46	75.20 ±0.39	83.80 ±0.51	64.90 ±0.57	68.80 ±0.48	74.00 ±0.55	45.80 ±0.52	50.50 ±0.46	54.80 ±0.58
MoE-Adapters	68.80 ±0.48	76.50 ±0.42	85.00 ±0.47	65.80 ±0.54	69.60 ±0.47	74.70 ±0.53	46.70 ±0.48	51.40 ±0.43	55.70 ±0.51
Pi-CCA	69.10 ±0.42	76.80 ±0.36	85.40 ±0.44	66.40 ±0.51	70.40 ±0.42	75.50 ±0.49	47.40 ±0.46	52.20 ±0.41	56.60 ±0.47
InfLoRA	66.90 ±0.53	74.60 ±0.45	83.70 ±0.52	66.50 ±0.59	68.60 ±0.49	74.40 ±0.57	—	—	—
SD-LoRA	67.00 ±0.58	72.90 ±0.52	79.60 ±0.61	64.90 ±0.63	67.30 ±0.55	72.60 ±0.62	—	—	—
KeepLoRA	69.30 ±0.41	77.50 ±0.37	86.00 ±0.43	67.40 ±0.48	70.80 ±0.41	75.90 ±0.47	—	—	—
NuSA-CL	68.70 ±0.43	75.50 ±0.40	83.30 ±0.48	68.00 ±0.46	70.40 ±0.39	75.60 ±0.45	—	—	—
BFP (ours)	**70.10** ±0.35	**78.20** ±0.32	**86.70** ±0.38	**70.06** ±0.44	**71.60** ±0.37	**77.00** ±0.43	**48.90** ±0.41	**53.70** ±0.36	**58.10** ±0.43

Each cell shows the mean above its ± sample SD. All entries are common-protocol evaluations. InfLoRA, SD-LoRA, KeepLoRA, and NuSA-CL are evaluated on MTIL only. Auxiliary data, teachers, and expanding experts follow Appendix B.1.3.

C.2.4 BACKBONE AND BOUNDARY-DENSITY COMPARISONS

Table 9: **Retrieval across backbones and stream granularities.** Final eight-domain macro $R@1$ (%, ↑). All models use OpenAI weights at 224×224. Eight-stage training uses 1,000 updates/domain; each 32-stage variant uses 250 updates/group, preserving 8,000 updates in total. "Contiguous" and "interleaved" follow Appendix B.1.1; galleries and domain weights are identical.

	8 stages		32, contiguous		32, interleaved	
Method	I2T	T2I	I2T	T2I	I2T	T2I
ViT-B/16						
Sequential LoRA (Hu et al., 2022)	33.12	32.46	32.95	32.31	32.51	31.89
C-CLIP (Liu et al., 2025)	40.11	37.74	39.36	37.12	38.68	36.45
Pi-CCA (Zhang et al., 2026)	40.94	38.52	40.28	37.96	39.71	37.34
BFP (ours)	**42.05**	**39.58**	**41.82**	**39.41**	**41.46**	**39.06**
ViT-B/32						
Sequential LoRA	31.46	30.84	31.29	30.72	30.91	30.33
C-CLIP	38.36	35.92	37.61	35.30	36.94	34.67
Pi-CCA	39.02	36.61	38.43	36.08	37.86	35.51
BFP (ours)	**40.18**	**37.71**	**39.97**	**37.53**	**39.62**	**37.21**
ViT-L/14						
Sequential LoRA	35.27	34.82	35.06	34.61	34.69	34.25
C-CLIP	44.02	41.37	43.23	40.69	42.51	40.01
Pi-CCA	44.76	42.14	44.11	41.54	43.54	40.92
BFP (ours)	**45.93**	**43.08**	**45.68**	**42.86**	**45.29**	**42.51**

These are common-protocol evaluations with equal update counts. More frequent boundaries add probing and SVD costs to the fixed update budget.

C.3 DIRECTION-LEVEL PROTECTION-UTILITY STATISTICS

The paired measurements and utility region are defined in Appendix B.6.

Table 10 reports region membership, gain concentration, mean overlap scores, and pooled Pearson correlations. Under growing protection, directions in $\mathcal{U}$ have more than twice the mean overlap score of those outside it (0.698 versus 0.319). BFP has a smaller utility-region population and retains a positive overlap–gain correlation ($r = 0.779$).

Table 10: **Direction-level utility statistics.** Ten paired runs at $t = 7$; $\overline{\mathcal{U}}$ denotes directions outside the utility region.

Statistic	Growing protection	BFP
Total directions	4,431	160
Directions in $\mathcal{U}$	727	8
Fraction in $\mathcal{U}$ (%)	16.4	5.0
Positive-gain share $S_{\mathcal{U}}$ (%)	47.3	21.7
Mean q_i in $\mathcal{U}$	0.698	0.586
Mean q_i in $\overline{\mathcal{U}}$	0.319	0.310
Pearson $r(q_i, G_i)$	0.883	0.779
Pearson $r(\log(1+\lambda_i), G_i)$	0.047	−0.348

C.4 MECHANISM CONTROLS AND PROTECTION OVERHEAD

The supplementary comparisons use only the seven-domain ViT-B/16 diagnostic stream, ten paired seeds, and the fixed configuration in Table 4. Sketch remains held out from parent training. The growing-protection and BFP entries in Table 11 reproduce the corresponding full-protection results in Table 3. Scores are percentages, paired differences are percentage points, and performance cells use mean ± sample SD. In Tables 11–12, boldface and underlining mark the highest and second-highest means within each metric and, for the history table, each checkpoint. These are descriptive comparisons; the primary analysis uses the seven hypotheses in Appendix B.3.2.

C.4.1 PROTECTION CAPACITY AND REFRESHED DIRECTIONS

The parent controls are defined in Appendix B.2.4. Table 11 compares each parent's full-protection diagnostic branch at $t = 7$; the table reports each control's realized rank.

Table 11: **Fixed-coordinate parent comparison.** Late-checkpoint Sketch adaptation and retention after $K = 200$ updates; k_7 is the minimum–maximum realized protection rank over ten runs.

Parent	k_7	$\text{AUC}_7 \uparrow$	$R_7 \uparrow$	$Z_7 \uparrow$
No protection	0	**60.28** ±1.63	47.71 ±0.90	62.39 ±0.49
Growing protection	438–447	50.42 ±1.73	**51.36** ±0.82	<u>63.12</u> ±0.44
Frozen first-task basis	16	58.96 ±1.58	49.14 ±0.85	62.91 ±0.46
Fixed random basis	16	<u>60.13</u> ±1.61	47.88 ±0.88	62.44 ±0.48
BFP (ours)	16	57.86 ±1.55	<u>50.98</u> ±0.78	**63.34** ±0.41

C.4.2 EARLY AND MIDDLE CHECKPOINTS

The primary full/half branches define the history contrast. Table 12 supplies their absolute performance at $t = 1, 4$, alongside the late-checkpoint values in Table 3. Historical retention averages the t domains encountered by that checkpoint, so both the model state and historical evaluation set change across checkpoints.

Table 12: **Absolute performance along the diagnostic history.** Superscript $-$ denotes half release. All six scores are percentages (↑).

Parent	t	AUC_t	AUC_t^-	R_t	R_t^-	Z_t	Z_t^-
Growing protection	1	<u>58.74</u> ±1.61	**59.99** ±1.67	**54.72** ±1.01	**54.60** ±1.03	<u>64.08</u> ±0.40	<u>64.04</u> ±0.41
	4	<u>54.21</u> ±1.68	<u>58.17</u> ±1.74	**52.63** ±0.88	**52.42** ±0.90	<u>63.56</u> ±0.42	<u>63.48</u> ±0.43
BFP (ours)	1	**59.13** ±1.50	<u>59.57</u> ±1.53	<u>54.61</u> ±0.99	<u>54.54</u> ±1.00	**64.12** ±0.39	**64.09** ±0.40
	4	**58.39** ±1.53	**59.01** ±1.57	<u>52.41</u> ±0.85	<u>52.30</u> ±0.86	**63.71** ±0.40	**63.65** ±0.41

C.4.3 STEP LENGTH, ORIENTATION, AND OPTIMIZER STATE

Table 13 compares each norm-replayed control with the original full-protection branch at $t = 7$. The reset contrast is reset-half minus reset-full; both start with zero moments and counters at the same learned parameters. SDs use paired run differences.

Table 13: **Late-checkpoint diagnostic controls.** Mean ± sample SD of paired effects (pp); positive effects indicate higher scores. e_{norm} is the maximum absolute displacement-norm mismatch over all completed control steps and runs. Failures count infeasible runs out of ten.

Control / paired contrast	δAUC	δR	δZ	e_{norm}	Failures
(a) Growing-protection parent					
Norm-replayed half release	$+5.18$ ±0.92	-0.24 ±0.31	-0.10 ±0.24	1.7×10^{-7}	0/10
Norm-replayed random deletion	$+6.34$ ±1.06	-0.71 ±0.42	-0.27 ±0.30	2.0×10^{-7}	0/10
Norm-replayed Haar basis	$+7.41$ ±1.18	-1.36 ±0.53	-0.54 ±0.35	2.2×10^{-7}	0/10
Reset half − reset full	$+6.52$ ±1.09	-0.30 ±0.38	-0.13 ±0.28	—	—
(b) BFP (ours) parent					
Norm-replayed half release	$+0.57$ ±0.35	-0.12 ±0.27	-0.05 ±0.20	1.5×10^{-7}	0/10
Norm-replayed random deletion	$+0.83$ ±0.43	-0.27 ±0.31	-0.11 ±0.22	1.6×10^{-7}	0/10
Norm-replayed Haar basis	$+1.19$ ±0.51	-0.55 ±0.37	-0.24 ±0.26	1.7×10^{-7}	0/10
Reset half − reset full	$+0.73$ ±0.46	-0.16 ±0.32	-0.08 ±0.23	—	—

Norm-replayed controls retain $\lfloor k_7/2 \rfloor$ directions and share reference norms. Deletion samples columns of U_7; Haar replaces their orientation. Reset branches are unrescaled (dashes: inapplicable). The ten-run comparison requires all controls to be feasible; failures remain in the reported run count.

C.4.4 PROTECTION RANK, STORAGE, AND COMPONENT TIME

Table 14 reports protection costs along both parent trajectories. The measurement definitions are given in Appendix B.2.5.

Table 14: **Protection overhead along the parent stream.** Rank is the run-wise minimum–maximum; storage and timing cells are mean ± sample SD.

Parent	t	Rank k_t	State (MiB)	Proj. (ms)	Refresh (s)
Growing protection	1	61–64	8.86 ±0.15	0.006 ±0.001	1.84 ±0.07
	4	249–255	35.51 ±0.29	0.081 ±0.007	2.71 ±0.12
	7	438–447	62.31 ±0.42	0.123 ±0.009	4.08 ±0.18
BFP (ours)	1	16	2.250 ±0.000	0.006 ±0.001	1.85 ±0.07
	4	16	2.250 ±0.000	0.043 ±0.004	1.93 ±0.08
	7	16	2.250 ±0.000	0.044 ±0.004	1.94 ±0.08

D PROOFS

D.1 SPECTRAL AND PROJECTION GEOMETRY

Proposition 3 (Optimal residual sensitivity)**.** *For $C \succeq 0$ and $0 \le r < d$, the optimum in* (6) *is $\lambda_{r+1}(C)$. Its leading $\min(r, \operatorname{rank} C)$ positive-eigenvalue directions attain it.*

Proof. Let E_{r+1} be a leading $(r+1)$-dimensional eigenspace. For a feasible rank-k basis, $\dim(\ker U^\top \cap E_{r+1}) \ge (d-k) + (r+1) - d \ge 1$. A unit vector in this intersection has Rayleigh quotient at least $\lambda_{r+1}(C)$. Protecting the leading r directions attains this bound when $\operatorname{rank} C > r$. Otherwise, protecting all positive-eigenvalue directions leaves zero residual sensitivity. □

For this basis, $\delta^\top C \delta \le \lambda_{r+1}(C)\|\delta\|_2^2$ whenever $U^\top \delta = 0$. The same basis maximizes $\operatorname{tr}(U^\top C U)$, and hence minimizes total residual sensitivity $\operatorname{tr}(P_U C P_U) = \operatorname{tr}(C) - \operatorname{tr}(U^\top C U)$. Both statements concern the matrix C.

The constrained least-squares problem in (8) has Lagrangian $\frac{1}{2}\|v-u\|_2^2 + \alpha^\top U^\top v$. Stationarity and $U^\top U = I$ give $\alpha = U^\top u$ and the unique solution $v = P_U u$. For a fixed minibatch with locally L-Lipschitz gradient and SGD proposal $u = g$, the descent lemma reduces to

$$\ell(\theta - \eta P_U g; \mathcal{B}) \le \ell(\theta; \mathcal{B}) - \eta \left(1 - \frac{L\eta}{2}\right) \|P_U g\|_2^2. \tag{33}$$

This uses $g^\top P_U g = \|P_U g\|_2^2$. For $0 < \eta < 2/L$ and $P_U g \neq 0$, the projected SGD step decreases the minibatch loss. An adaptive proposal can point uphill after projection: with $g = (1,1)^\top$, $u = (2,-1)^\top$, and $U = (1,0)^\top$, $g^\top u = 1$ but $g^\top P_U u = -1$.

For nested bases, $R = UU^\top - U^-(U^-)^\top$ is an orthogonal projector, $P_{U^-} = P_U + R$, and $P_U R = 0$. Pythagoras proves (9). For a fixed nonzero g and a Haar-random rank-k basis, rotational symmetry gives $\mathbb{E}[\|P_U g\|_2^2/\|g\|_2^2] = 1 - k/d$. A learned basis can intercept the entire gradient: a nonzero $g \in \operatorname{span}(U)$ has no surviving component even when $d - k$ is large.

D.2 ACCUMULATED COMPRESSION ERROR

We prove Proposition 2 using the stored statistic and residuals defined in Appendix A.3.1.

Proof. Subtract the two averaging recurrences to obtain $D_n = (n-1)D_{n-1}/n + E_n$. Multiplying by n and telescoping gives the identity. Each discarded eigenvalue is nonnegative, with $\|E_j\|_2 = \lambda_{r+1}(M_j)$. The spectral bound follows from the triangle inequality, and the trace identity from $\operatorname{tr}(C_n^{\text{all}}) = 1$. □

D.3 FINITE-STEP MODEL AND NOTATION

Throughout this appendix, vector norms are Euclidean and matrix norms are spectral unless a subscript indicates otherwise. All matrices are finite-dimensional and all identities are in exact arithmetic. The theory conditions on a single checkpoint and its fixed nested protection spaces. We use Γ for a constrained-optimum gap and $\mathcal{G}_K$ for a finite-step loss difference, reserving Δ_t for the measured Recall@1 intervention.

We analyze the shared-checkpoint intervention in Section 4.2. Output-gradient projection and its kernel-regime analysis provide related foundations (Farajtabar et al., 2020; Bennani et al., 2020); here the object is the signed, finite-step loss difference between two nested constraints at the same checkpoint. The results concern deterministic projected gradient descent on a quadratic diagnostic objective, with a controlled smooth extension in Appendix D.10. We analyze objective-loss differences; the experimental diagnostic measures Recall@1 AUC through (3).

Write displacements from the common checkpoint as $\delta = \theta - \theta_0$. Let the orthonormal columns of V span $\ker U^\top$, and let those of W span the released directions, so that $Z = [V, W]$ spans $\ker(U^-)^\top$. Thus $VV^\top = P_U$ and $WW^\top = P_{U^-} - P_U$. We first take the original free space and the released

subspace to have positive dimensions; Appendix D.12 treats empty spaces. On the larger allowed space, consider

$$f(\delta) = f_0 - b^\top \delta + \tfrac{1}{2}\delta^\top \mathsf{H}\delta, \qquad b = -\nabla f(0), \qquad M = Z^\top \mathsf{H} Z, \qquad 0 < \mu I \preceq M \preceq LI. \tag{34}$$

Only the curvature on $\mathrm{span}(Z)$ is restricted. This includes fixed-feature least squares with a positive-definite restricted Gram matrix. All branches start at $\delta_0 = 0$, use the same $0 < \eta \le L^{-1}$, and keep their spaces fixed:

$$\delta^Q_{k+1} = \delta^Q_k - \eta Q Q^\top \nabla f(\delta^Q_k), \quad Q \in \{V, Z\}, \qquad \mathcal{G}_K = f(\delta^V_K) - f(\delta^Z_K). \tag{35}$$

A positive $\mathcal{G}_K$ is a loss-based release benefit. The empirical PIPL criterion additionally requires historical and pretrained retention. Throughout, $t^0 = 1$ also when $t = 0$. Define the block geometry and its Schur complement by

$$\begin{gathered} A = V^\top \mathsf{H} V, \quad B = V^\top \mathsf{H} W, \quad D = W^\top \mathsf{H} W, \qquad M = \begin{bmatrix} A & B \\ B^\top & D \end{bmatrix}, \\ a = V^\top b, \quad c = W^\top b, \quad h = \begin{bmatrix} a \\ c \end{bmatrix}, \quad S = D - B^\top A^{-1} B, \quad \zeta = c - B^\top A^{-1} a. \end{gathered} \tag{36}$$

The vector ζ is the remaining released-coordinate descent signal after optimizing the original free coordinates. It differs from the initial released gradient c when curvature couples the two spaces.

For $Q \in \{V, Z\}$, put $T_Q = Q^\top \mathsf{H} Q$, $b_Q = Q^\top b$, and define

$$\mathcal{D}_Q = \tfrac{1}{2} b_Q^\top T_Q^{-1} b_Q, \qquad \mathcal{E}_Q(K) = \tfrac{1}{2} b_Q^\top T_Q^{-1} (I - \eta T_Q)^{2K} b_Q, \qquad f^*_Q = \min_{\delta \in \mathrm{span}(Q)} f(\delta). \tag{37}$$

Define the scalar optimization filter, the uncoupled release benefit, and coupling magnitude by

$$\begin{gathered} \varphi_K(\lambda) = \frac{1 - (1 - \eta\lambda)^{2K}}{\lambda} = \eta \sum_{j=0}^{2K-1} (1 - \eta\lambda)^j, \qquad \mathcal{J}_K = \tfrac{1}{2} c^\top \varphi_K(D) c, \qquad \beta = \|B\|_2, \\ q = 1 - \eta\mu, \qquad C_{1,K} = \eta^2 \sum_{j=1}^{2K-1} j q^{j-1}, \qquad C_{2,K} = \eta^3 \sum_{j=2}^{2K-1} \binom{j}{2} q^{j-2}. \end{gathered} \tag{38}$$

Empty sums are zero; powers with exponent zero are one, including at $q = 0$.

For the released block, write $\mu_D = \lambda_{\min}(D)$ and $L_D = \lambda_{\max}(D)$. These definitions apply throughout the appendix; the initial free and released spaces are assumed nonempty until Appendix D.12.

D.4 Restricted geometry and elimination of free coordinates

Lemma 4 (Schur geometry with inherited curvature bounds)**.** *Suppose $\mu I \preceq M \preceq LI$ with $\mu > 0$. Then A, D are positive definite, and the Schur complement satisfies $\mu I \preceq S \preceq LI$. For every released coordinate y, minimizing over the original free coordinate x gives*

$$\begin{aligned} f(Vx + Wy) = f_0 - \tfrac{1}{2} a^\top A^{-1} a + \tfrac{1}{2}\big(x - A^{-1}(a - By)\big)^\top A \big(x - A^{-1}(a - By)\big) \\ + \tfrac{1}{2} y^\top S y - \zeta^\top y. \end{aligned} \tag{39}$$

Consequently, the released minimizer has $y_* = S^{-1}\zeta$ and $x_* = A^{-1}(a - By_*)$, and

$$\mathcal{D}_Z = \mathcal{D}_V + \tfrac{1}{2}\zeta^\top S^{-1}\zeta, \qquad \frac{\|\zeta\|^2}{2L} \le \Gamma = \tfrac{1}{2}\zeta^\top S^{-1}\zeta \le \frac{\|\zeta\|^2}{2\mu}. \tag{40}$$

Proof. Testing the quadratic form of M on vectors $(x, 0)$ or $(0, y)$ yields, respectively,

$$\mu\|x\|^2 \le \begin{bmatrix} x \\ 0 \end{bmatrix}^\top M \begin{bmatrix} x \\ 0 \end{bmatrix} = x^\top A x \le L\|x\|^2, \qquad \mu\|y\|^2 \le y^\top D y \le L\|y\|^2. \tag{41}$$

Thus A^{-1} exists. Completing a square in a homogeneous quadratic form gives the exact variational representation

$$\begin{bmatrix} x \\ y \end{bmatrix}^\top M \begin{bmatrix} x \\ y \end{bmatrix} = (x + A^{-1}By)^\top A(x + A^{-1}By) + y^\top Sy,$$
$$y^\top Sy = \min_x \begin{bmatrix} x \\ y \end{bmatrix}^\top M \begin{bmatrix} x \\ y \end{bmatrix}, \qquad x_{\min} = -A^{-1}By. \tag{42}$$

Eliminating the original free coordinates preserves the lower curvature bound:

$$y^\top Sy = \min_x \begin{bmatrix} x \\ y \end{bmatrix}^\top M \begin{bmatrix} x \\ y \end{bmatrix} \geq \min_x \mu(\|x\|^2 + \|y\|^2) = \mu\|y\|^2,$$
$$y^\top Sy \leq \begin{bmatrix} 0 \\ y \end{bmatrix}^\top M \begin{bmatrix} 0 \\ y \end{bmatrix} = y^\top Dy \leq L\|y\|^2. \tag{43}$$

This proves the stated bounds on S and its invertibility. Expanding the right-hand side of (39) recovers $f_0 - a^\top x - c^\top y + \frac{1}{2}x^\top Ax + x^\top By + \frac{1}{2}y^\top Dy$: the cross term in $\frac{1}{2}(a - By)^\top A^{-1}(a - By)$ changes the linear coefficient of y from c to ζ, and the quadratic term changes D to S. Since A and S are positive definite, the two remaining squares have unique minimizers. More explicitly,

$$\begin{aligned} \min_x f(Vx + Wy) &= f_0 - \tfrac{1}{2}a^\top A^{-1}a + \tfrac{1}{2}y^\top Sy - \zeta^\top y \\ &= f_0 - \mathcal{D}_V - \tfrac{1}{2}\zeta^\top S^{-1}\zeta + \tfrac{1}{2}(y - S^{-1}\zeta)^\top S(y - S^{-1}\zeta) \\ &\geq f_0 - \mathcal{D}_V - \tfrac{1}{2}\zeta^\top S^{-1}\zeta, \end{aligned} \tag{44}$$

with equality exactly at $y = S^{-1}\zeta$. The restricted Z-objective also has minimum $f_0 - \frac{1}{2}h^\top M^{-1}h$, proving the value identity. Finally, diagonalizing S with eigenvalues $s_i \in [\mu, L]$ and coordinates ζ_i gives the full comparison

$$\frac{1}{2L}\sum_i \zeta_i^2 \leq \frac{1}{2}\sum_i \frac{\zeta_i^2}{s_i} = \frac{1}{2}\zeta^\top S^{-1}\zeta \leq \frac{1}{2\mu}\sum_i \zeta_i^2. \tag{45}$$

The endpoint constants are sharp given $\mu, L, \|\zeta\|$: choose $B = 0$ and support $\zeta = c$ on an eigendirection of D with eigenvalue L or μ. Additional unused eigendirections can realize both prescribed spectral endpoints. □

D.5 Projected dynamics and sharp spectral residuals

Lemma 5 (Restricted gradient descent as a spectral filter)**.** *Let Q have orthonormal columns, $T = Q^\top \mathsf{H} Q \succ 0$, $v = Q^\top b$, and $0 < \eta \leq 1/\lambda_{\max}(T)$. The recursion* (35) *remains in* $\mathrm{span}(Q)$ *and has the exact solution*

$$\delta_K^Q = Qz_K, \qquad z_K = T^{-1}[I - (I - \eta T)^K]v, \qquad z_K - z_* = -(I - \eta T)^K T^{-1}v, \quad z_* = T^{-1}v. \tag{46}$$

If $Te_i = \lambda_i e_i$ and $v_i = e_i^\top v$, then

$$f(\delta_K^Q) - f_Q^* = \tfrac{1}{2}\sum_i \frac{v_i^2}{\lambda_i}(1 - \eta\lambda_i)^{2K},$$
$$f_0 - f(\delta_K^Q) = \tfrac{1}{2}\sum_i \frac{v_i^2}{\lambda_i}[1 - (1 - \eta\lambda_i)^{2K}] = \tfrac{1}{2}v^\top \varphi_K(T)v. \tag{47}$$

Proof. The initial displacement is zero. Every update is in $\mathrm{span}(Q)$, so induction shows that $\delta_k^Q = Qz_k$. Multiplying the update by $Q^\top$ and using $Q^\top Q = I$ yields

$$z_{k+1} = (I - \eta T)z_k + \eta v, \qquad z_{k+1} - T^{-1}v = (I - \eta T)(z_k - T^{-1}v). \tag{48}$$

Iterating from $z_0 = 0$ proves the error formula. Equivalently, the finite geometric sum telescopes without an approximation:

$$\begin{aligned} z_K &= \eta \sum_{j=0}^{K-1} (I - \eta T)^j v, \\ Tz_K &= \sum_{j=0}^{K-1} \big[(I - \eta T)^j - (I - \eta T)^{j+1}\big] v = [I - (I - \eta T)^K] v. \end{aligned} \tag{49}$$

The restricted objective completes to

$$f(Qz) = f_0 - \tfrac{1}{2} v^\top T^{-1} v + \tfrac{1}{2}(z - T^{-1}v)^\top T (z - T^{-1}v). \tag{50}$$

Since every polynomial in T commutes with T^{-1}, inserting (46) gives

$$\begin{aligned} f(\delta_K^Q) - f_Q^* &= \tfrac{1}{2} v^\top T^{-1} (I - \eta T)^K T (I - \eta T)^K T^{-1} v \\ &= \tfrac{1}{2} v^\top T^{-1} (I - \eta T)^{2K} v = \tfrac{1}{2} \sum_i \frac{v_i^2}{\lambda_i} (1 - \eta \lambda_i)^{2K}. \end{aligned} \tag{51}$$

Subtracting this expression from $f_0 - f_Q^* = \frac{1}{2} v^\top T^{-1} v$ proves the optimization-benefit identity. The polynomial representation of φ_K follows by telescoping as in (49), now with $2K$ terms. □

Lemma 6 (Endpoint residual envelopes and their equality cases)**.** *With the preceding notation, let $\mathcal{D} = \frac{1}{2} v^\top T^{-1} v$, $\mu_T = \lambda_{\min}(T)$, and $L_T = \lambda_{\max}(T)$. For every $K \geq 0$,*

$$(1 - \eta L_T)^{2K} \mathcal{D} \leq f(\delta_K^Q) - f_Q^* \leq (1 - \eta \mu_T)^{2K} \mathcal{D}. \tag{52}$$

Both constants are attainable. If $v \neq 0$, the bounds can be strengthened by replacing μ_T, L_T with the smallest and largest eigenvalues whose eigenspaces have nonzero projection of v.

Proof. For every eigenvalue, the step-size restriction ensures

$$0 \leq 1 - \eta L_T \leq 1 - \eta \lambda_i \leq 1 - \eta \mu_T < 1. \tag{53}$$

The weights $v_i^2/(2\lambda_i)$ are nonnegative. Therefore the comparison is termwise and preserves all signs:

$$\begin{aligned} (1 - \eta L_T)^{2K} \mathcal{D} &= \frac{1}{2} \sum_i \frac{v_i^2}{\lambda_i} (1 - \eta L_T)^{2K} \\ &\leq \frac{1}{2} \sum_i \frac{v_i^2}{\lambda_i} (1 - \eta \lambda_i)^{2K} = f(\delta_K^Q) - f_Q^* \\ &\leq \frac{1}{2} \sum_i \frac{v_i^2}{\lambda_i} (1 - \eta \mu_T)^{2K} = (1 - \eta \mu_T)^{2K} \mathcal{D}. \end{aligned} \tag{54}$$

For $K \geq 1$ and nonzero v, equality at an endpoint holds when v lies in its eigenspace; conversely, strict monotonicity of $\lambda \mapsto (1 - \eta\lambda)^{2K}$ on the relevant interval forces this support condition. At $K = 0$ both bounds are equalities for every v. Removing terms with zero weight proves the active-spectrum refinement. If $v = 0$, all iterates, decrements, and residuals vanish, so no active eigenvalue needs to be defined. □

D.6 Exact finite-step decomposition and sufficient budget

Theorem 7 (Exact loss-based PIPL and finite-horizon certificate)**.** *Under* (34)–(35)*, for every integer $K \geq 0$,*

$$\begin{aligned} f(\delta_K^Q) = f_Q^* + \mathcal{E}_Q(K), \qquad f_Q^* = f_0 - \mathcal{D}_Q, \\ \Gamma := f_V^* - f_Z^* = \tfrac{1}{2} \zeta^\top S^{-1} \zeta \geq 0, \quad \mathcal{G}_K = \Gamma + \mathcal{E}_V(K) - \mathcal{E}_Z(K). \end{aligned} \tag{55}$$

Let $\mu_Q = \lambda_{\min}(T_Q)$, $L_Q = \lambda_{\max}(T_Q)$, $\rho_Q = 1 - \eta\mu_Q$, and $\sigma_Q = 1 - \eta L_Q$. Then

$$\begin{aligned} \Gamma + \sigma_V^{2K} \mathcal{D}_V - \rho_Z^{2K} (\mathcal{D}_V + \Gamma) &\leq \mathcal{G}_K \\ &\leq \Gamma + \rho_V^{2K} \mathcal{D}_V - \sigma_Z^{2K} (\mathcal{D}_V + \Gamma). \end{aligned} \tag{56}$$

Each constituent residual envelope is sharp. If $\Gamma > 0$, the condition $\rho_Z^{2K} < \Gamma / \mathcal{D}_Z$ suffices for $\mathcal{G}_K > 0$. Moreover, $\mu I \preceq S \preceq LI$, and hence $\|\zeta\|^2/(2L) \leq \Gamma \leq \|\zeta\|^2/(2\mu)$.

The identity separates the persistent price of protection, Γ, from the two optimization transients, whose difference determines finite-step loss ordering. Appendix D.11 gives a two-dimensional example with $\mathcal{G}_1 = -7/800$ despite a stable step size and $\Gamma > 0$. Equation (56) combines bounds that are sharp for each residual separately.

Complete proof of Theorem 7. Lemma 5 applied to $Q = V$ and $Q = Z$ gives both objective decompositions. Lemma 4 gives $f_V^* - f_Z^* = \Gamma$ and $\mathcal{D}_Z = \mathcal{D}_V + \Gamma$. Subtraction is exact:

$$\begin{aligned} \mathcal{G}_K &= [f_V^* + \mathcal{E}_V(K)] - [f_Z^* + \mathcal{E}_Z(K)] \\ &= \Gamma + \mathcal{E}_V(K) - \mathcal{E}_Z(K). \end{aligned} \tag{57}$$

The lower and upper bounds require opposite endpoint choices for the subtracted residual. Keeping this sign reversal explicit,

$$\begin{aligned} \mathcal{G}_K &= \Gamma + \mathcal{E}_V(K) - \mathcal{E}_Z(K) \\ &\geq \Gamma + \sigma_V^{2K}\mathcal{D}_V - \mathcal{E}_Z(K) \\ &\geq \Gamma + \sigma_V^{2K}\mathcal{D}_V - \rho_Z^{2K}\mathcal{D}_Z \\ &= \Gamma + \sigma_V^{2K}\mathcal{D}_V - \rho_Z^{2K}(\mathcal{D}_V + \Gamma), \\ \mathcal{G}_K &\leq \Gamma + \rho_V^{2K}\mathcal{D}_V - \mathcal{E}_Z(K) \\ &\leq \Gamma + \rho_V^{2K}\mathcal{D}_V - \sigma_Z^{2K}\mathcal{D}_Z \\ &= \Gamma + \rho_V^{2K}\mathcal{D}_V - \sigma_Z^{2K}(\mathcal{D}_V + \Gamma). \end{aligned} \tag{58}$$

These are precisely (56). To certify a strictly positive gap, one may use the sharper displayed lower endpoint directly. A simpler sufficient condition follows by retaining only nonnegativity of the original residual:

$$\mathcal{G}_K = \Gamma + \mathcal{E}_V(K) - \mathcal{E}_Z(K) \geq \Gamma - \mathcal{E}_Z(K) \geq \Gamma - \rho_Z^{2K}\mathcal{D}_Z > 0. \tag{59}$$

The final strict inequality is exactly the condition in the theorem. The curvature and inverse bounds for S were proved in Lemma 4. Attainability of each residual endpoint follows from Lemma 6. □

If $\Gamma > 0$ and $0 < \rho_Z < 1$, an integer budget that satisfies the simple strict certificate is

$$K_{\mathrm{cert}} = \left\lfloor \frac{\log(\mathcal{D}_Z/\Gamma)}{-2\log\rho_Z} \right\rfloor + 1, \qquad K \geq K_{\mathrm{cert}} \implies \mathcal{G}_K > 0. \tag{60}$$

Indeed, $\mathcal{D}_Z = \mathcal{D}_V + \Gamma \geq \Gamma > 0$, so the logarithm is defined and nonnegative. The floor plus one makes the inequality strict even when the ratio is an integer. If $\rho_Z = 0$, one step reaches the restricted quadratic optimum and any $K \geq 1$ suffices. The certificate applies when $\Gamma > 0$ and gives a sufficient budget; profitable release may occur earlier.

D.7 Proof of the main finite-step recovery theorem

Lemma 8 (Monotone finite-step filters)**.** *For $K \geq 1$ and $0 < \lambda \leq L$ with $0 < \eta \leq 1/L$, $\varphi_K(\lambda) > 0$ is strictly decreasing in λ. The function $\psi_K(\lambda) = 1 - (1-\eta\lambda)^{2K}$ is increasing in λ, and both filters are nondecreasing in K.*

Proof. The finite geometric identity gives $\varphi_K(\lambda) = \eta\sum_{j=0}^{2K-1}(1-\eta\lambda)^j$. Every summand is nonnegative, the $j = 0$ summand is strictly positive, and for $\lambda_1 \leq \lambda_2$ in the stated interval,

$$\begin{aligned} 0 \leq 1 - \eta\lambda_2 \leq 1 - \eta\lambda_1 < 1, \\ \varphi_K(\lambda_2) = \eta\sum_{j=0}^{2K-1}(1-\eta\lambda_2)^j \leq \eta\sum_{j=0}^{2K-1}(1-\eta\lambda_1)^j = \varphi_K(\lambda_1), \\ \psi_K(\lambda_1) = 1 - (1-\eta\lambda_1)^{2K} \leq 1 - (1-\eta\lambda_2)^{2K} = \psi_K(\lambda_2). \end{aligned} \tag{61}$$

If $\lambda_1 < \lambda_2$, the $j = 1$ term makes the middle inequality strict. For horizon monotonicity, write $t = 1 - \eta\lambda \in [0, 1)$; then

$$\varphi_{K+1}(\lambda) - \varphi_K(\lambda) = \eta(t^{2K} + t^{2K+1}) \geq 0, \qquad \psi_{K+1}(\lambda) - \psi_K(\lambda) = t^{2K}(1-t^2) \geq 0. \tag{62}$$

This proves all claims, including the endpoint $t = 0$. □

Complete proof of Theorem 1. When $B = 0$, M is block diagonal. Lemma 5 then gives identical original-free coordinates in the two branches. Subtracting their exact optimization benefits cancels those coordinates:

$$\begin{aligned}\mathcal{G}_K &= \tfrac{1}{2}h^\top \varphi_K(M)h - \tfrac{1}{2}a^\top \varphi_K(A)a \\ &= \tfrac{1}{2}\left[a^\top \varphi_K(A)a + c^\top \varphi_K(D)c\right] - \tfrac{1}{2}a^\top \varphi_K(A)a \\ &= \tfrac{1}{2}c^\top \varphi_K(D)c = \mathcal{J}_K.\end{aligned} \tag{63}$$

If $Du_i = d_i u_i$ and $c_i = u_i^\top c$, Lemma 8 yields two distinct, useful comparisons:

$$\begin{aligned}\tfrac{1}{2}\varphi_K(L_D)\|c\|^2 = \tfrac{1}{2}\sum_i c_i^2 \varphi_K(L_D) &\le \tfrac{1}{2}\sum_i c_i^2 \varphi_K(d_i) = \mathcal{J}_K \\ &\le \tfrac{1}{2}\sum_i c_i^2 \varphi_K(\mu_D) = \tfrac{1}{2}\varphi_K(\mu_D)\|c\|^2, \\ \psi_K(\mu_D)\Gamma = \tfrac{1}{2}\sum_i \frac{c_i^2}{d_i}\psi_K(\mu_D) &\le \tfrac{1}{2}\sum_i \frac{c_i^2}{d_i}\psi_K(d_i) = \mathcal{J}_K \\ &\le \tfrac{1}{2}\sum_i \frac{c_i^2}{d_i}\psi_K(L_D) = \psi_K(L_D)\Gamma.\end{aligned} \tag{64}$$

Notice that the eigenvalue endpoints reverse between the two comparisons: one fixes $\|c\|^2$, while the other fixes the energy $\Gamma = \frac{1}{2}c^\top D^{-1}c$. Both are sharp by taking c in the relevant extremal eigenspace. Their intersection can be used when both summaries are available. Since $\varphi_K(D) = D^{-1}[I - (I - \eta D)^{2K}]$, these identities and inequalities are exactly (12). For $K \ge 1$, the eigenvalues of $\varphi_K(D)$ are strictly positive, so $\mathcal{G}_K > 0$ if and only if $c \neq 0$; each decay factor tends to zero, giving $\mathcal{G}_K \to \Gamma$. This proves Theorem 1 and the invariant-space part of Theorem 9. □

For isotropic curvature $\mathsf{H} = I$ and $b = \theta_\star - \theta_0$, Theorem 9 specializes to

$$\mathcal{G}_K = \tfrac{1}{2}[1 - (1 - \eta)^{2K}]\|(P_{U^-} - P_U)(\theta_\star - \theta_0)\|^2. \tag{65}$$

Orthonormality gives $\|c\| = \|WW^\top b\|$, so (65) follows directly from the exact invariant-space formula. Thus released task signal, curvature, and the optimization budget jointly determine recovery.

D.8 Curvature coupling and second-order correction

Theorem 9 (Sharp invariant-space gain and coupling correction)**.** *Under Theorem 7, let $K \ge 1$, $\mu_D = \lambda_{\min}(D)$, and $L_D = \lambda_{\max}(D)$. Then*

$$\begin{aligned}|\mathcal{G}_K - \mathcal{J}_K| &\le \Xi_K, \\ \Xi_K = \min\left\{\tfrac{1}{2}\beta C_{1,K}(\|a\|^2 + \|c\|^2),\quad \beta C_{1,K}\|a\|\|c\| + \tfrac{1}{2}\beta^2 C_{2,K}(\|a\|^2 + \|c\|^2)\right\}&, \\ \tfrac{1}{2}\varphi_K(L_D)\|c\|^2 - \Xi_K \le \mathcal{G}_K \le \tfrac{1}{2}\varphi_K(\mu_D)\|c\|^2 + \Xi_K&.\end{aligned} \tag{66}$$

If $B = 0$, then $\mathcal{G}_K = \mathcal{J}_K \ge 0$ exactly; both gradient-norm endpoints in (66) *are attainable. In this case $\Gamma = \frac{1}{2}c^\top D^{-1}c$ and the additional sharp bounds are*

$$[1 - (1 - \eta\mu_D)^{2K}]\Gamma \le \mathcal{G}_K \le [1 - (1 - \eta L_D)^{2K}]\Gamma. \tag{67}$$

The first-order coupling coefficient in the second term defining Ξ_K is attainable at $K = 1$.

The condition $\frac{1}{2}\varphi_K(L_D)\|c\|^2 > \Xi_K$ certifies a finite-step gain beyond exact curvature invariance. The proof retains an off-diagonal first-order term before bounding the second-order remainder; this gives the sharper dependence on $\|a\|\|c\|$.

Lemma 10 (Two-level telescoping for contractions)**.** *Let N, N_0 be square matrices with $\|N\|, \|N_0\| \le q$, let $F = N - N_0$, and suppose $0 \le q < 1$. For an integer $j \ge 1$, define*

$$T_j = \sum_{\ell=0}^{j-1} N_0^\ell F N_0^{j-1-\ell}, \qquad R_j = N^j - N_0^j - T_j. \tag{68}$$

Then

$$\|N^j - N_0^j\| \le jq^{j-1}\|F\|, \qquad \|R_j\| \le \binom{j}{2} q^{j-2}\|F\|^2 \quad (j \ge 2), \qquad R_1 = 0. \tag{69}$$

Proof. No commutation is required. Consecutive terms cancel in the ordered identity

$$\begin{aligned} N^j - N_0^j &= \sum_{\ell=0}^{j-1} \big(N_0^\ell N^{j-\ell} - N_0^{\ell+1} N^{j-1-\ell}\big) \\ &= \sum_{\ell=0}^{j-1} N_0^\ell F N^{j-1-\ell}. \end{aligned} \tag{70}$$

Submultiplicativity gives the first chain:

$$\begin{aligned} \|N^j - N_0^j\| &\le \sum_{\ell=0}^{j-1} \|N_0^\ell F N^{j-1-\ell}\| \\ &\le \sum_{\ell=0}^{j-1} \|N_0\|^\ell \|F\| \|N\|^{j-1-\ell} \\ &\le \sum_{\ell=0}^{j-1} q^\ell \|F\| q^{j-1-\ell} = j q^{j-1} \|F\|. \end{aligned} \tag{71}$$

Subtracting the first-order term while preserving multiplication order gives

$$R_j = \sum_{\ell=0}^{j-1} N_0^\ell F (N^{j-1-\ell} - N_0^{j-1-\ell}). \tag{72}$$

The $\ell = j-1$ term vanishes. Applying the first bound to each remaining power yields

$$\begin{aligned} \|R_j\| &\le \sum_{\ell=0}^{j-2} \|N_0\|^\ell \|F\| \|N^{j-1-\ell} - N_0^{j-1-\ell}\| \\ &\le \sum_{\ell=0}^{j-2} q^\ell \|F\| (j-1-\ell) q^{j-2-\ell} \|F\| \\ &= q^{j-2} \|F\|^2 \sum_{\ell=0}^{j-2} (j-1-\ell) = \binom{j}{2} q^{j-2} \|F\|^2. \end{aligned} \tag{73}$$

For $j = 1$, the identity directly gives $T_1 = F$ and $R_1 = 0$. The stated formulas are polynomial at $q = 0$, with nonnegative exponents in their respective ranges. □

Complete proof of Theorem 9. The invariant case and its sharp scalar bounds were proved in Appendix D.7. For general B, put

$$M_0 = \begin{bmatrix} A & 0 \\ 0 & D \end{bmatrix}, \quad E = M - M_0 = \begin{bmatrix} 0 & B \\ B^\top & 0 \end{bmatrix}, \quad N = I - \eta M, \quad N_0 = I - \eta M_0, \quad F = -\eta E. \tag{74}$$

Principal-submatrix bounds imply $\mu I \preceq M_0 \preceq LI$. Since $E^2 = \mathrm{diag}(BB^\top, B^\top B)$ and E is symmetric, its norm is $\beta = \|B\|$. Therefore

$$\begin{aligned} \|N\| &= \max_{\lambda \in \mathrm{spec}(M)} (1 - \eta\lambda) \le 1 - \eta\mu = q, \\ \|N_0\| &\le 1 - \eta\mu = q, \qquad \|F\| = \eta \|E\| = \eta\beta. \end{aligned} \tag{75}$$

The exact benefit formula in Lemma 5, and block-diagonal cancellation for M_0, give

$$\begin{aligned} \mathcal{G}_K - \mathcal{J}_K &= \tfrac{1}{2} h^\top [\varphi_K(M) - \varphi_K(M_0)] h \\ &= \tfrac{\eta}{2} \sum_{j=1}^{2K-1} h^\top (N^j - N_0^j) h. \end{aligned} \tag{76}$$

The $j = 0$ terms cancel exactly. Lemma 10 gives a first, dimension-free bound:

$$\begin{aligned}|\mathcal{G}_K - \mathcal{J}_K| &\le \tfrac{\eta}{2} \sum_{j=1}^{2K-1} |h^\top (N^j - N_0^j) h| \\ &\le \tfrac{\eta}{2}\|h\|^2 \sum_{j=1}^{2K-1} \|N^j - N_0^j\| \\ &\le \tfrac{\eta}{2}\|h\|^2 \sum_{j=1}^{2K-1} j q^{j-1}\eta\beta = \tfrac{1}{2}\beta C_{1,K}(\|a\|^2 + \|c\|^2).\end{aligned} \tag{77}$$

To improve the dependence on the two gradient blocks, retain the first-order structure. Write $N_A = I - \eta A$ and $N_D = I - \eta D$. As N_0 is block diagonal and F is off diagonal, T_j has zero diagonal blocks and upper-right block $-\eta \sum_{\ell=0}^{j-1} N_A^\ell B N_D^{j-1-\ell}$. The lower-left block is its transpose after reindexing. Thus

$$h^\top T_j h = -2\eta \sum_{\ell=0}^{j-1} a^\top N_A^\ell B N_D^{j-1-\ell} c. \tag{78}$$

Applying Cauchy–Schwarz before replacing the vector blocks by a single norm preserves a useful cancellation:

$$\begin{aligned}|h^\top T_j h| &\le 2\eta \sum_{\ell=0}^{j-1} |a^\top N_A^\ell B N_D^{j-1-\ell} c| \\ &\le 2\eta \sum_{\ell=0}^{j-1} \|N_A^\ell a\| \, \|B N_D^{j-1-\ell} c\| \\ &\le 2\eta\beta\|a\|\|c\| \sum_{\ell=0}^{j-1} \|N_A\|^\ell \|N_D\|^{j-1-\ell} \\ &\le 2j\eta\beta q^{j-1}\|a\|\|c\|.\end{aligned} \tag{79}$$

The first-order term vanishes if either gradient block vanishes. The remaining contribution is second order in β. Combining (73) and (79), with $R_1 = 0$, gives

$$\begin{aligned}|\mathcal{G}_K - \mathcal{J}_K| &\le \tfrac{\eta}{2} \sum_{j=1}^{2K-1} \big(|h^\top T_j h| + |h^\top R_j h|\big) \\ &\le \tfrac{\eta}{2} \sum_{j=1}^{2K-1} |h^\top T_j h| + \tfrac{\eta}{2}\|h\|^2 \sum_{j=2}^{2K-1} \|R_j\| \\ &\le \eta^2\beta\|a\|\|c\| \sum_{j=1}^{2K-1} j q^{j-1} + \tfrac{\eta^3\beta^2}{2}\|h\|^2 \sum_{j=2}^{2K-1} \binom{j}{2} q^{j-2} \\ &= \beta C_{1,K}\|a\|\|c\| + \tfrac{1}{2}\beta^2 C_{2,K}(\|a\|^2 + \|c\|^2).\end{aligned} \tag{80}$$

Both estimates hold simultaneously, so their minimum is valid. The invariant-space endpoint comparison bounds $\mathcal{J}_K$ for any $D \succ 0$, irrespective of whether the actual B is zero. Hence

$$\begin{aligned}\mathcal{G}_K &\ge \mathcal{J}_K - |\mathcal{G}_K - \mathcal{J}_K| \ge \mathcal{J}_K - \Xi_K \ge \tfrac{1}{2}\varphi_K(L_D)\|c\|^2 - \Xi_K, \\ \mathcal{G}_K &\le \mathcal{J}_K + |\mathcal{G}_K - \mathcal{J}_K| \le \mathcal{J}_K + \Xi_K \le \tfrac{1}{2}\varphi_K(\mu_D)\|c\|^2 + \Xi_K.\end{aligned} \tag{81}$$

For first-step sharpness, $\varphi_1(T) = 2\eta I - \eta^2 T$, so the correction is exactly

$$\mathcal{G}_1 - \mathcal{J}_1 = -\eta^2 a^\top B c, \qquad C_{1,1} = \eta^2, \quad C_{2,1} = 0. \tag{82}$$

Choose nonzero a, c along a left and right singular-vector pair of B attaining β, with matching or opposite signs. Then the absolute correction is $\eta^2\beta\|a\|\|c\|$. A positive-definite realization is $A = D = \alpha I$ with $0 < \beta < \alpha$, taking the step size at most $1/(\alpha + \beta)$. The refined bound is then attained; by $2\|a\|\|c\| \le \|a\|^2 + \|c\|^2$ it is also the smaller term in Ξ_1. This establishes the stated sharpness and completes the proof. □

Proposition 11 (Signed correction and horizon-uniform constants)**.** *With $N_A = I - \eta A$ and $N_D = I - \eta D$, define*

$$\chi_K = \eta^2 \sum_{j=1}^{2K-1} \sum_{\ell=0}^{j-1} a^\top N_A^\ell B N_D^{j-1-\ell} c. \tag{83}$$

Then the more informative, signed approximation satisfies

$$|\mathcal{G}_K - (\mathcal{J}_K - \chi_K)| \le \tfrac{1}{2}\beta^2 C_{2,K}\|h\|^2, \qquad \begin{aligned} C_{1,K} &\le \min\{\eta^2 K(2K-1), \mu^{-2}\}, \\ C_{2,K} &\le \min\{\eta^3 \tbinom{2K}{3}, \mu^{-3}\}. \end{aligned} \tag{84}$$

Proof. Equations (76) and (78) give the exact identity

$$\mathcal{G}_K - \mathcal{J}_K = -\chi_K + \tfrac{\eta}{2} \sum_{j=2}^{2K-1} h^\top R_j h. \tag{85}$$

Bounding only the remainder by (73) proves the first claim. The signed term captures whether curvature coupling helps or hinders release. A refined first-order norm bound uses the separate contractions $q_A = \|N_A\|$ and $q_D = \|N_D\|$:

$$|\chi_K| \le \eta^2 \beta \|a\|\|c\| \sum_{j=1}^{2K-1} \sum_{\ell=0}^{j-1} q_A^\ell q_D^{j-1-\ell} \le \beta C_{1,K}\|a\|\|c\|. \tag{86}$$

For the constants, $q \in [0, 1)$ and finite geometric sums give two complementary bounds:

$$\begin{aligned} C_{1,K} &= \eta^2 \sum_{j=1}^{2K-1} j q^{j-1} \le \eta^2 \sum_{j=1}^{2K-1} j = \eta^2 K(2K-1), \\ C_{1,K} &\le \eta^2 \sum_{j=1}^{\infty} j q^{j-1} = \frac{\eta^2}{(1-q)^2} = \mu^{-2}, \\ C_{2,K} &= \eta^3 \sum_{j=2}^{2K-1} \binom{j}{2} q^{j-2} \le \eta^3 \sum_{j=2}^{2K-1} \binom{j}{2} = \eta^3 \binom{2K}{3}, \\ C_{2,K} &\le \eta^3 \sum_{j=2}^{\infty} \binom{j}{2} q^{j-2} = \frac{\eta^3}{(1-q)^3} = \mu^{-3}. \end{aligned} \tag{87}$$

For completeness, the first infinite series equals the product of two geometric series: the coefficient of q^m in $(\sum_{i\ge 0} q^i)^2$ is $m+1$. The second equals the product of three such series, whose coefficient is $\binom{m+2}{2}$. These products are justified by absolute convergence. The finite binomial sum follows by telescoping $\binom{j+1}{3} - \binom{j}{3} = \binom{j}{2}$. Thus the correction constants remain bounded as the horizon grows. □

D.9 Loss-area identities on a fixed evaluation grid

Corollary 12 (Nonnegative weighted loss-area comparison)**.** *Let $w_0, \dots, w_K \ge 0$ sum to one and be fixed independently of the branch outcomes. Define $\overline{\mathcal{G}}_w = \sum_{k=0}^K w_k \mathcal{G}_k$ and $\overline{\varphi}_w(\lambda) = \sum_{k=0}^K w_k \varphi_k(\lambda)$, with $\varphi_0 = 0$. Then*

$$\overline{\mathcal{G}}_w = \Gamma + \sum_{k=0}^{K} w_k \mathcal{E}_V(k) - \sum_{k=0}^{K} w_k \mathcal{E}_Z(k). \tag{88}$$

The coupling comparison in (66) *applies after replacing $\mathcal{G}_K$ by $\overline{\mathcal{G}}_w$, $\mathcal{J}_K$ by $\frac{1}{2}c^\top \overline{\varphi}_w(D)c$, φ_K by $\overline{\varphi}_w$, and Ξ_K by $\sum_k w_k \Xi_k$, where $\Xi_0 = 0$. In particular, $B = 0$ implies $\overline{\mathcal{G}}_w \ge 0$.*

Proof. Summing (55) and using $\sum_k w_k = 1$ gives (88), including any weight at $k = 0$, where $\mathcal{G}_0 = 0$ and $\mathcal{D}_Z - \mathcal{D}_V = \Gamma$. Nonnegativity of the weights then preserves the direction of each bound:

$$\begin{aligned}\left|\overline{\mathcal{G}}_w - \tfrac{1}{2}c^\top \overline{\varphi}_w(D)c\right| &= \left|\sum_{k=0}^{K} w_k(\mathcal{G}_k - \mathcal{J}_k)\right| \\ &\le \sum_{k=0}^{K} w_k|\mathcal{G}_k - \mathcal{J}_k| \le \sum_{k=0}^{K} w_k \Xi_k.\end{aligned} \tag{89}$$

When $B = 0$, every correction vanishes and every scalar filter is nonnegative. Summing the endpoint bounds likewise proves sharp invariant-space bounds for the weighted filter: a common endpoint eigenvector attains equality simultaneously at all grid points. □

Normalized trapezoidal integration of a loss curve on a fixed grid has nonnegative weights and is therefore covered. For the simple post-update mean, $w_k = 1/K$ for $1 \le k \le K$, let $t = 1 - \eta\lambda$. Direct summation gives

$$\overline{\varphi}_K(\lambda) = \frac{1}{K\lambda}\sum_{k=1}^{K}(1 - t^{2k}) = \frac{1}{\lambda}\left[1 - \frac{t^2(1 - t^{2K})}{K(1 - t^2)}\right], \qquad 0 \le t < 1. \tag{90}$$

The denominator is positive because $\lambda, \eta > 0$; at $t = 0$ the displayed fraction is zero. The areas integrate the diagnostic objective loss over the specified grid.

D.10 Extension to smooth, nonquadratic losses

We now bound the approximation error when a smooth diagnostic loss is replaced by its checkpoint Taylor quadratic. Let $\mathcal{F}$ be a twice continuously differentiable loss in ambient displacement coordinates, and let its checkpoint Taylor quadratic be $f(\delta) = \mathcal{F}(0) + \nabla\mathcal{F}(0)^\top\delta + \frac{1}{2}\delta^\top\nabla^2\mathcal{F}(0)\delta$. Thus $b = -\nabla\mathcal{F}(0)$ and $\mathsf{H} = \nabla^2\mathcal{F}(0)$ in the preceding results.

Fix one orthonormal free-space basis Q. Set $F_Q(y) = \mathcal{F}(Qy)$, $v = Q^\top b$, and $T = Q^\top \mathsf{H} Q$. Assume there is an open convex coordinate region Ω_Q containing the actual and Taylor-quadratic iterates through step K, including zero, such that

$$\begin{aligned}&0 < \mu_Q^{\mathrm{loc}} I \preceq \nabla^2 F_Q(y) \preceq L_Q^{\mathrm{loc}} I \quad (y \in \Omega_Q), \\ &\|\nabla^2 F_Q(y) - \nabla^2 F_Q(z)\| \le \mathcal{M}_Q\|y - z\| \quad (y, z \in \Omega_Q), \qquad 0 < \eta \le (L_Q^{\mathrm{loc}})^{-1}.\end{aligned} \tag{91}$$

The same η is used for both branches when their release gap is compared. The comparison applies while both trajectories remain in the region Ω_Q satisfying (91). Define actual and quadratic coordinate trajectories by

$$y_0 = z_0 = 0, \qquad y_{k+1} = y_k - \eta\nabla F_Q(y_k), \qquad z_{k+1} = z_k + \eta(v - Tz_k). \tag{92}$$

For the following results suppress the subscript Q on the local constants, and put

$$\begin{aligned}m = \lambda_{\min}(T), \quad \ell = \lambda_{\max}(T), \quad q = 1 - \eta\mu^{\mathrm{loc}}, \quad q_T = 1 - \eta m, \quad G = \|v\|, \\ R_k = \frac{G}{\mu^{\mathrm{loc}}}(1 - q^k), \qquad Z_k = \|z_k\|, \qquad a_k = \|(I - \eta T)^k v\|.\end{aligned} \tag{93}$$

Here a_k is a nonnegative scalar residual, distinct from the block vector a in (36).

Lemma 13 (Contraction, trajectory radii, and Taylor remainder)**.** *Under* (91)*, $\|y_k\| \le R_k$, $Z_k \le R_k$, and $a_k \le q_T^k G \le q^k G$. For any $y \in \Omega_Q$, define $e(y) = \nabla F_Q(y) + v - Ty$ and $\tau(y) = F_Q(y) - f(Qy)$. Then*

$$\|e(y)\| \le \tfrac{\mathcal{M}}{2}\|y\|^2, \qquad |\tau(y)| \le \tfrac{\mathcal{M}}{6}\|y\|^3. \tag{94}$$

Proof. Let $\mathcal{T}(y) = y - \eta\nabla F_Q(y)$. For $u, w \in \Omega_Q$, convexity of the region and the fundamental theorem of calculus give

$$\begin{aligned}\mathcal{T}(u) - \mathcal{T}(w) &= (I - \eta\overline{H}_{u,w})(u - w), \\ \overline{H}_{u,w} &= \int_0^1 \nabla^2 F_Q(w + s(u - w))\,\mathrm{d}s, \qquad \mu^{\mathrm{loc}} I \preceq \overline{H}_{u,w} \preceq L^{\mathrm{loc}} I.\end{aligned} \tag{95}$$

The integral is symmetric and preserves the quadratic-form bounds. Since the step size makes every eigenvalue of $I - \eta\overline{H}_{u,w}$ nonnegative,

$$\begin{aligned}\|\mathcal{T}(u) - \mathcal{T}(w)\| &\le \|I - \eta\overline{H}_{u,w}\|\|u - w\| \\ &= \max_i |1 - \eta\lambda_i(\overline{H}_{u,w})|\|u - w\| \\ &\le (1 - \eta\mu^{\mathrm{loc}})\|u - w\| = q\|u - w\|.\end{aligned} \tag{96}$$

The first increment has length ηG. For $k \ge 1$, comparing consecutive iterates and then telescoping their increments yields

$$\begin{aligned}\|y_{k+1} - y_k\| &\le q\|y_k - y_{k-1}\| \le \cdots \le q^k\|y_1 - y_0\| = \eta G q^k, \\ \|y_k\| &\le \sum_{j=0}^{k-1} \|y_{j+1} - y_j\| \le \eta G \sum_{j=0}^{k-1} q^j = \frac{G}{\mu^{\mathrm{loc}}}(1 - q^k) = R_k.\end{aligned} \tag{97}$$

For the quadratic, let u_i be an orthonormal eigenbasis of T and put $v_i = u_i^\top v$. Diagonalization gives the exact iterate radius and residual:

$$\begin{aligned}Z_k^2 &= \sum_i v_i^2 \left[\frac{1 - (1 - \eta\lambda_i(T))^k}{\lambda_i(T)}\right]^2, \\ a_k^2 &= \sum_i v_i^2 (1 - \eta\lambda_i(T))^{2k}, \qquad \nabla_y f(Qz_k) = -(I - \eta T)^k v.\end{aligned} \tag{98}$$

For each eigenvalue, the nonnegative geometric sum is at most $\eta \sum_{j=0}^{k-1} q^j$ because $\lambda_i(T) \ge \mu^{\mathrm{loc}}$. Summing its square with weights v_i^2 proves $Z_k \le R_k$. The residual bound follows similarly with $q_T \le q$.

Finally, the segment from zero to y lies in Ω_Q. Subtracting the first-order Taylor gradient and integrating its Hessian variation gives

$$\begin{aligned}e(y) &= \int_0^1 [\nabla^2 F_Q(sy) - T]y \,\mathrm{d}s, \\ \|e(y)\| &\le \int_0^1 \|\nabla^2 F_Q(sy) - T\| \,\|y\| \,\mathrm{d}s \\ &\le \int_0^1 \mathcal{M}s\|y\|^2 \,\mathrm{d}s = \tfrac{\mathcal{M}}{2}\|y\|^2.\end{aligned} \tag{99}$$

A second integration yields the scalar remainder, with its full constant:

$$\begin{aligned}\tau(y) &= \int_0^1 (1 - s)y^\top [\nabla^2 F_Q(sy) - T]y \,\mathrm{d}s, \\ |\tau(y)| &\le \int_0^1 (1 - s)\|\nabla^2 F_Q(sy) - T\| \,\|y\|^2 \,\mathrm{d}s \\ &\le \mathcal{M}\|y\|^3 \int_0^1 s(1 - s) \,\mathrm{d}s = \tfrac{\mathcal{M}}{6}\|y\|^3.\end{aligned} \tag{100}$$

The constants $1/2$ and $1/6$ are attained by a scalar cubic perturbation on a sufficiently small interval with positive curvature. □

Lemma 14 (Two complementary iterate-error recursions)**.** *Define $\epsilon_0 = 0$ and, using only the quadratic path and known constants,*

$$\epsilon_{k+1} = \min\left\{q_T\epsilon_k + \tfrac{\eta\mathcal{M}}{2}R_k^2, \quad q\epsilon_k + \tfrac{\eta\mathcal{M}}{2}Z_k^2\right\}. \tag{101}$$

Then $\|y_k - z_k\| \le \epsilon_k$ and $\|y_k\| \le \rho_k := \min\{R_k, Z_k + \epsilon_k\}$. In particular, $\epsilon_1 = 0$. For $K \ge 1$, a less refined closed-form envelope is

$$\begin{aligned}\epsilon_K &\le \frac{\eta\mathcal{M}G^2}{2(\mu^{\mathrm{loc}})^2} \sum_{t=0}^{K-1} q^{K-1-t}(1 - q^t)^2 \\ &= \frac{\mathcal{M}G^2}{2(\mu^{\mathrm{loc}})^3}\left[(1 + q^{K-1})(1 - q^K) - 2K(1 - q)q^{K-1}\right].\end{aligned} \tag{102}$$

Proof. Let $d_k = y_k - z_k$. Exact subtraction of (92) gives $d_{k+1} = (I - \eta T)d_k - \eta e(y_k)$. Thus

$$\begin{aligned}\|d_{k+1}\| &\le \|I - \eta T\|\|d_k\| + \eta\|e(y_k)\| \\ &\le q_T\|d_k\| + \tfrac{\eta\mathcal{M}}{2}\|y_k\|^2 \\ &\le q_T\|d_k\| + \tfrac{\eta\mathcal{M}}{2}R_k^2.\end{aligned} \tag{103}$$

A different decomposition uses contraction of the nonlinear map, evaluated on the quadratic path:

$$\begin{aligned}\|d_{k+1}\| &= \|\mathcal{T}(y_k) - [z_k + \eta(v - Tz_k)]\| \\ &\le \|\mathcal{T}(y_k) - \mathcal{T}(z_k)\| + \|\mathcal{T}(z_k) - [z_k + \eta(v - Tz_k)]\| \\ &\le q\|d_k\| + \eta\|e(z_k)\| \\ &\le q\|d_k\| + \tfrac{\eta\mathcal{M}}{2}Z_k^2.\end{aligned} \tag{104}$$

Both affine upper bounds are nondecreasing in $\|d_k\|$. Starting from $d_0 = 0$, induction therefore permits their minimum at every step and proves (101). The triangle inequality and the independent radius bound imply

$$\|y_k\| \le \|z_k\| + \|d_k\| \le Z_k + \epsilon_k, \qquad \|y_k\| \le R_k, \qquad \|y_k\| \le \rho_k. \tag{105}$$

Since $R_0 = Z_0 = 0$, the first actual and quadratic updates agree and $\epsilon_1 = 0$.

For an explicit scalar envelope, use $q_T \le q$ and $Z_t \le R_t$, then unwind the recursion:

$$\begin{aligned}\epsilon_K &\le \tfrac{\eta\mathcal{M}}{2}\sum_{t=0}^{K-1} q^{K-1-t}R_t^2 \\ &= \frac{\eta\mathcal{M}G^2}{2(\mu^{\mathrm{loc}})^2}\sum_{t=0}^{K-1} q^{K-1-t}(1-q^t)^2 \\ &\le \tfrac{\eta\mathcal{M}}{2}R_{K-1}^2\sum_{t=0}^{K-1} q^{K-1-t} = \frac{\mathcal{M}}{2\mu^{\mathrm{loc}}}(1-q^K)R_{K-1}^2.\end{aligned} \tag{106}$$

To derive the equality in (102), expand the square in the finite sum:

$$\begin{aligned}\sum_{t=0}^{K-1} q^{K-1-t}(1-q^t)^2 &= \sum_{t=0}^{K-1} q^{K-1-t} - 2\sum_{t=0}^{K-1} q^{K-1} + \sum_{t=0}^{K-1} q^{K-1+t} \\ &= \frac{1-q^K}{1-q} - 2Kq^{K-1} + q^{K-1}\frac{1-q^K}{1-q}.\end{aligned} \tag{107}$$

Using $1 - q = \eta\mu^{\mathrm{loc}}$ proves the formula. The finite sum is numerically preferable when cancellation is substantial. At $q = 0$, the equality is interpreted as a polynomial identity with $0^0 = 1$; in fact uniform equal curvature then makes the nonlinear loss quadratic on the region. The transfer bound below uses the sharper recursive estimate. □

Theorem 15 (Asymmetric finite-step transfer from the checkpoint quadratic)**.** *Under* (91)*, compute* ϵ_K, ρ_K, a_K *as above and put*

$$\begin{aligned}t_K &= \min\{\epsilon_K, a_K/m\}, \qquad \Psi_K = a_K t_K - \tfrac{m}{2}t_K^2, \\ \mathcal{B}_Q^-(K) &= \Psi_K + \tfrac{\mathcal{M}}{6}\rho_K^3, \qquad \mathcal{B}_Q^+(K) = a_K\epsilon_K + \tfrac{\ell}{2}\epsilon_K^2 + \tfrac{\mathcal{M}}{6}\rho_K^3.\end{aligned} \tag{108}$$

Then, for each branch,

$$-\mathcal{B}_Q^-(K) \le \mathcal{F}(Qy_K) - f(Qz_K) \le \mathcal{B}_Q^+(K). \tag{109}$$

Apply the construction separately to $Q = V$ and $Q = Z$. Let $\mathcal{G}_K^{\mathcal{F}}$ be their actual nonlinear loss difference and $\mathcal{G}_K^f$ the exact quadratic gap of Theorem 7. Then

$$\mathcal{G}_K^f - \mathcal{B}_V^-(K) - \mathcal{B}_Z^+(K) \le \mathcal{G}_K^{\mathcal{F}} \le \mathcal{G}_K^f + \mathcal{B}_V^+(K) + \mathcal{B}_Z^-(K). \tag{110}$$

In particular, $\mathcal{G}_K^f > \mathcal{B}_V^-(K) + \mathcal{B}_Z^+(K)$ certifies a positive nonlinear release benefit.

Proof. Expand the quadratic at its own iterate. The error $d_K = y_K - z_K$ stays in the same free-coordinate space, whose quadratic residual has norm a_K. Exactly,

$$\begin{aligned}\mathcal{F}(Qy_K) - f(Qz_K) &= \tau(y_K) + f(Qy_K) - f(Qz_K) \\ &= \tau(y_K) + (-v + Tz_K)^\top d_K + \tfrac{1}{2} d_K^\top T d_K.\end{aligned} \tag{111}$$

For the upper error, the quadratic term has the same sign as its spectral upper bound:

$$\begin{aligned}\mathcal{F}(Qy_K) - f(Qz_K) &\le \tfrac{\mathcal{M}}{6}\|y_K\|^3 + \|-v + Tz_K\|\|d_K\| + \tfrac{\ell}{2}\|d_K\|^2 \\ &\le \tfrac{\mathcal{M}}{6}\rho_K^3 + a_K\|d_K\| + \tfrac{\ell}{2}\|d_K\|^2 \\ &\le \tfrac{\mathcal{M}}{6}\rho_K^3 + a_K\epsilon_K + \tfrac{\ell}{2}\epsilon_K^2 = \mathcal{B}_Q^+(K).\end{aligned} \tag{112}$$

For the lower error, retaining positive curvature gives a tighter result than discarding the quadratic term. Put $s = \|d_K\| \in [0, \epsilon_K]$. Then

$$\begin{aligned}\mathcal{F}(Qy_K) - f(Qz_K) &\ge -\tfrac{\mathcal{M}}{6}\|y_K\|^3 - \|-v + Tz_K\|\|d_K\| + \tfrac{m}{2}\|d_K\|^2 \\ &\ge -\tfrac{\mathcal{M}}{6}\rho_K^3 - a_K s + \tfrac{m}{2}s^2 \\ &\ge -\tfrac{\mathcal{M}}{6}\rho_K^3 + \min_{0\le u\le\epsilon_K}\{-a_K u + \tfrac{m}{2}u^2\} \\ &= -\tfrac{\mathcal{M}}{6}\rho_K^3 - a_K t_K + \tfrac{m}{2}t_K^2 = -\mathcal{B}_Q^-(K).\end{aligned} \tag{113}$$

The minimizer of the scalar convex quadratic is the unconstrained minimizer a_K/m clipped to $[0, \epsilon_K]$, which proves the last equality. In particular,

$$0 \le \Psi_K = \max_{0\le u\le\epsilon_K}(a_K u - \tfrac{m}{2}u^2) \le \min\{a_K\epsilon_K, a_K^2/(2m)\}. \tag{114}$$

Thus this lower error never exceeds the simpler bound obtained by taking the absolute value of the linear term. The scalar optimization is sharp when only (a_K, m, ϵ_K) are retained: a one-dimensional quadratic with residual of magnitude a_K, curvature m, and displacement opposite that residual attains it.

Finally, write the actual loss difference as the quadratic difference plus its two branch errors, denoted $\varepsilon_V^{\mathrm{loss}}$ and $\varepsilon_Z^{\mathrm{loss}}$. Equation (109) gives

$$\begin{aligned}\mathcal{G}_K^{\mathcal{F}} &= \mathcal{G}_K^f + \varepsilon_V^{\mathrm{loss}} - \varepsilon_Z^{\mathrm{loss}} \\ &\ge \mathcal{G}_K^f - \mathcal{B}_V^-(K) - \mathcal{B}_Z^+(K), \\ \mathcal{G}_K^{\mathcal{F}} &\le \mathcal{G}_K^f + \mathcal{B}_V^+(K) + \mathcal{B}_Z^-(K).\end{aligned} \tag{115}$$

This proves the two-sided comparison and its strict positivity certificate. □

The error depends on the decaying *free-space* residual a_K, even when the protected component of the ambient gradient remains large. When $\mathcal{M} = 0$, every transfer error is zero and the quadratic theory is recovered exactly. At $K = 1$ the iterate error is zero, although the scalar Taylor remainder at that common first iterate can be nonzero. A readily interpretable local nonlinearity ratio is $\mathcal{M}G/(\mu^{\mathrm{loc}})^2$: the convolution estimate bounds the coordinate discrepancy, uniformly over horizons for which the stated assumptions hold, by $\mathcal{M}G^2/[2(\mu^{\mathrm{loc}})^3]$. The transfer certificate is useful only when its explicit errors are smaller than the quadratic release margin. The interval combines separate bounds for the two branch errors.

D.11 Counterexamples to unconditional finite-step improvement

Proposition 16 (Release can hurt before it helps)**.** *There exists a strongly convex two-dimensional quadratic, nested allowed spaces, and $\eta \le 1/L$ for which $\Gamma > 0$ but $\mathcal{G}_1 < 0$.*

Proof. Take $V = e_1$, $W = e_2$, and

$$\mathsf{H} = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 1/10 \end{bmatrix}, \qquad f(x,y) = \tfrac{1}{2}(x^2 + xy + y^2) - x - y/10, \qquad \eta = \tfrac{1}{2}. \tag{116}$$

The Hessian eigenvalues are $1/2$ and $3/2$, so $\eta = 1/2 \le 2/3 = 1/L$. The first protected and released displacements are $(1/2, 0)$ and $(1/2, 1/20)$. Substitution, with no limiting approximation, gives

$$\begin{aligned} f(\delta_1^V) &= \tfrac{1}{2}(1/4) - 1/2 = -3/8 = -300/800, \\ f(\delta_1^Z) &= \tfrac{1}{2}(1/4 + 1/40 + 1/400) - 1/2 - 1/200 = -293/800, \\ \mathcal{G}_1 &= -7/800 < 0. \end{aligned} \tag{117}$$

Nevertheless $A = D = 1$, $B = 1/2$, $S = 3/4$, and $\zeta = 1/10 - 1/2 = -2/5$, so

$$\Gamma = \tfrac{1}{2}\,\frac{(-2/5)^2}{3/4} = \frac{8}{75} > 0, \qquad \|P_{U^-} b\|^2 - \|P_U b\|^2 = \frac{1}{100} > 0. \tag{118}$$

The first-step signed correction explains the reversal exactly: $\mathcal{J}_1 = 3/800$ and $\eta^2 a^\top B c = 10/800$, whose difference is $-7/800$. Here curvature coupling outweighs the recovered gradient component at the first step, despite the stable step size. □

Proposition 17 (Transient PIPL with a zero optimum gap)**.** *It is possible to have* $\Gamma = 0$*,* $\mathcal{G}_1 > 0$*, and* $\mathcal{G}_2 < 0$ *for the same nested pair and stable step size.*

Proof. Keep the same Hessian, spaces, and $\eta = 1/2$, but set $b = (1, 1/2)^\top = \mathsf{H} e_1$. Both constrained optima are e_1, so $\zeta = 0$ and $\Gamma = 0$. The protected initial optimality gap is $1/2$. The released Hessian eigenvectors $(1, 1)^\top/\sqrt{2}$ and $(1, -1)^\top/\sqrt{2}$ have eigenvalues $3/2$ and $1/2$, and their initial spectral energy weights are $3/8$ and $1/8$. Thus the exact finite-step gap is

$$\mathcal{G}_K = \tfrac{1}{2}(1/2)^{2K} - \tfrac{3}{8}(1/4)^{2K} - \tfrac{1}{8}(3/4)^{2K}. \tag{119}$$

At the first two horizons this gives

$$\begin{aligned} \mathcal{G}_1 &= \frac{1}{8} - \frac{3}{128} - \frac{9}{128} = \frac{1}{32} > 0, \\ \mathcal{G}_2 &= \frac{1}{32} - \frac{3}{2048} - \frac{81}{2048} = -\frac{5}{512} < 0, \qquad \lim_{K\to\infty} \mathcal{G}_K = 0. \end{aligned} \tag{120}$$

The residualized signal ζ characterizes the constrained-optimum gap. Curvature coupling can change the optimization transient even when the original free space already contains the common optimum. □

D.12 Degenerate spaces, coordinate invariance, and scope

If no directions are released, W has zero columns and $Z = V$: both trajectories coincide and every gap is zero. Quantities requiring eigenvalues of the empty matrix D are then unnecessary; the invariant and coupling statements are interpreted through the exact zero identity. If the original free space has zero dimension but W does not, the protected branch is fixed at the checkpoint, $\mathcal{D}_V = \mathcal{E}_V(K) = 0$, $S = D$, and the exact released benefit is $\frac{1}{2} c^\top \varphi_K(D) c$. There is no coupling matrix to bound. If both spaces are empty, neither branch moves. At $K = 0$ both branches start at zero, so $\mathcal{G}_0 = 0$; in spectral endpoint formulas $t^0 = 1$ also when $t = 0$.

A zero projected initial gradient in a branch makes its quadratic trajectory stationary. For the smooth deterministic recursion, $\nabla F_Q(0) = 0$ likewise fixes that branch at zero, and its radius and transfer errors vanish. For an empty allowed space, set $\mathcal{B}_Q^-(K) = \mathcal{B}_Q^+(K) = 0$ directly, without evaluating eigenvalues of an empty Hessian; the paired transfer statement remains valid. These stationary-point statements concern the momentum-free updates defined above.

All exact identities and bounds are invariant to orthonormal changes of coordinates *within* either of the two fixed spaces. To see this, let O_1, O_2 be orthogonal, replace V, W by VO_1, WO_2, and write primes for the resulting matrices. Then

$$\begin{aligned} &A' = O_1^\top A O_1, \quad B' = O_1^\top B O_2, \quad D' = O_2^\top D O_2, \quad a' = O_1^\top a, \quad c' = O_2^\top c, \\ &S' = O_2^\top S O_2, \qquad \zeta' = O_2^\top \zeta, \qquad M' = \operatorname{diag}(O_1, O_2)^\top M \operatorname{diag}(O_1, O_2). \end{aligned} \tag{121}$$

For any polynomial p, repeated cancellation of $OO^\top = I$ proves $p(O^\top TO) = O^\top p(T)O$; inverses obey the same identity because $(O^\top TO)(O^\top T^{-1}O) = I$. Consequently,

$$\begin{aligned} (\zeta')^\top (S')^{-1}\zeta' &= \zeta^\top S^{-1}\zeta, \\ (c')^\top \varphi_K(D')c' &= c^\top \varphi_K(D)c, \qquad \|B'\| = \|B\|, \\ (b'_Q)^\top (T'_Q)^{-1}(I-\eta T'_Q)^{2K} b'_Q &= b_Q^\top T_Q^{-1}(I-\eta T_Q)^{2K} b_Q. \end{aligned} \tag{122}$$

The norms are unchanged because orthogonal matrices preserve Euclidean lengths and the unit sphere in the variational definition of the spectral norm. The signed correction is also unchanged term by term under the same substitutions. The identities therefore depend on the two subspaces and are invariant to orthonormal basis choices within the fixed parameter coordinates.

The invariant-space endpoint bounds, first-step coupling coefficient, single-branch spectral residual constants, and scalar Taylor constants are sharp for their stated summaries. The coupled multi-step envelope and nonlinear two-branch transfer provide sufficient conditions for loss recovery. Throughout, the analysis uses deterministic projected gradient descent in fixed spaces, with positive restricted curvature and, for smooth losses, the local conditions in (91). Adaptation and retention in the VLM experiments are evaluated through the paired protocol.