# GESPU: RETAIN-ANCHORED SPECTRAL PROJECTION FOR BUDGETED LLM UNLEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Parameter-efficient machine unlearning requires purging requested information from a pretrained model while preserving retained utility within a strict deployment footprint. Existing low-rank adaptation methods decouple direction from capacity: updates are optimized in isotropic Euclidean space, while ranks are allocated uniformly or through post-hoc heuristics. Consequently, explicit negative deletion gradients inject extreme update energy that triggers catastrophic parameter collisions on retained tasks. We introduce Geometric Spectral Unlearning (GESPU), which formulates budgeted unlearning as a constrained optimization problem on a Kronecker-factored retain-curvature manifold. We prove that minimizing the resulting local surrogate admits a closed-form spectral solution via Eckart–Young singular-value truncation coupled with a discrete Multiple-Choice Knapsack problem, guaranteeing global surrogate optimality with certified descent under exact steps. Crucially, before retained loss gradients become active, the retain-metric factors act as an anisotropic filter, steering initial deletion updates away from sensitive retained pathways. A dual-timescale schedule optimizes low-dimensional core matrices between periodic subspace refreshes, exporting standard low-rank adapters with zero deployment overhead. Evaluated on TOFU and WMDP benchmarks across LLaMA-3 and Qwen2.5 models, GESPU establishes superior Pareto trade-offs over existing PEFT baselines, effectively suppressing target knowledge, preserving complex reasoning capabilities, and resisting knowledge recovery under benign relearning.

## 1 INTRODUCTION

Pretrained large language models (LLMs) acquire broad general capabilities alongside sensitive personal data, copyrighted content, and hazardous domain knowledge. Regulatory mandates such as the EU GDPR Article 17 and safety benchmarks like WMDP necessitate principled post-training mechanisms to purge targeted knowledge (Bourtoule et al., 2021; Eldan & Russinovich, 2023; Yao et al., 2024; Li et al., 2024). Because retraining from scratch is computationally prohibitive, machine unlearning addresses this challenge by selectively suppressing designated data influences while preserving utility on retained distributions. However, unlearning requires targeted, verifiable functional erasure: unlike modular capability composition or task-vector arithmetic (Shazeer et al., 2017; Hinton et al., 2015; Ilharco et al., 2023; Ma et al., 2026), merely routing away from components or subtracting weight vectors alters execution without providing verifiable functional deletion across diverse downstream prompts.

Selective knowledge suppression also introduces severe optimization challenges. Whereas standard fine-tuning induces gradual behavioral drift, deletion objectives such as DPO and Negative Preference Optimization (NPO) inject explicit negative forces to minimize target likelihoods (Rafailov et al., 2023; Zhang et al., 2024). Because representation pathways are shared across tasks (Ren & Sutherland, 2025), unconstrained negative updates frequently trigger destructive parameter collisions, causing sudden capability collapse on retained reasoning and coding competencies (Yang et al., 2025; Zheng et al., 2025). Preventing this degradation requires an optimization geometry grounded in retained-response sensitivity. Under deployment constraints, low-rank adaptation provides parameter efficiency (Hu et al., 2022), but isotropic Euclidean update norms fail to reflect directional predictive curvature: a minor displacement along sensitive directions severely degrades model utility, whereas uniform shrinkage needlessly impedes deletion. While dynamic low-rank



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.

methods such as AdaLoRA and SoRA adapt capacity (Zhang et al., 2023; Ding et al., 2023), their heuristic allocation scores remain decoupled from the curvature of retained tasks. To bridge this gap, we introduce **GeSPU** (*Geometric Spectral Unlearning*), which casts budgeted unlearning as a constrained optimization problem over a Kronecker-factored retain-curvature manifold (Martens & Grosse, 2015). By unifying predictive output stability with intermediate semantic anchoring, GeSPU maps the local quadratic surrogate into an isometric coordinate system where singular-value truncation and dynamic programming solve a Multiple-Choice Knapsack Problem (MCKP), globally allocating discrete ranks under a hard storage budget. Crucially, at initialization, where retention loss gradients vanish identically, GeSPU's curvature factors act as an anisotropic filter from step zero, proactively attenuating deletion components aligned with sensitive retain pathways and channeling update energy into dormant parameter directions before capability collapse can occur. Our contributions are summarized as follows:

- **Curvature-Anchored Budgeted Formulation**: We cast parameter-efficient unlearning as an augmented Lagrangian problem under discrete storage budgets, unifying predictive retention and latent semantic anchoring via a Kronecker-factored curvature metric.
- **Exact Spectral Optimality and Proactive Steering**: We prove that the non-convex budgeted retain surrogate admits an exact, globally optimal closed-form spectral solution via Eckart–Young truncation and MCKP dynamic programming, while analytically characterizing its proactive initialization filtering.
- **Scalable Deployment and Empirical Superiority**: We design a dual-timescale schedule optimizing low-dimensional core matrices under exact metric isometry, exporting standard LoRA adapters that establish superior Pareto trade-offs across TOFU and WMDP benchmarks while resisting benign relearning.

## 2 Related Work

**LLM Unlearning, Refusal, and Editing Formulations.** Machine unlearning methods modify parameters to excise designated corpora or hazardous behaviors (Eldan & Russinovich, 2023; Yao et al., 2024). While NPO and SimNPO stabilize negative preference objectives (Zhang et al., 2024; Fan et al., 2025), SCRUB enforces teacher–student retention and SSD dampens sensitive weights via Fisher importance (Kurmanji et al., 2023; Foster et al., 2024). Concurrently, model editing techniques (e.g., ROME, MEMIT) adjust localized factual associations through activation statistics (Meng et al., 2022; 2023), though localized insertion differs fundamentally from broader capability excision. Comprehensive benchmarks, including TOFU's fictitious profiles, WMDP's hazardous capabilities, and MUSE's multi-faceted privacy metrics (Maini et al., 2024; Li et al., 2024; Shi et al., 2025), reveal that superficial output suppression does not ensure authentic parameter erasure. In particular, Distinguishable Deletion decouples parameter-level erasure from inference-time refusal (Yang et al., 2026), as superficial refusal remains susceptible to recovery via benign relearning (Hu et al., 2025). To guarantee retention, SOUL leverages second-order optimization (Jia et al., 2024) and CEU incorporates primal–dual retention constraints (Entesari et al., 2025). GeSPU advances beyond these unconstrained or isolated formulations by solving joint subspace selection and discrete capacity allocation under operational parameter budgets.

**Parameter-Efficient Adaptation and Spectral Allocation.** Parameter-efficient fine-tuning reduces deployment overhead by factorizing weight updates or sharing projection matrices (Hu et al., 2022; Valipour et al., 2023; Kopiczko et al., 2024). To dynamically modulate capacity, AdaLoRA prunes singular components via gradient-weight importance scores, and SoRA applies proximal $\ell_1$ regularization to rank gates (Zhang et al., 2023; Ding et al., 2023); however, both optimize in isotropic Euclidean space without accounting for retained-task curvature. While Fisher information and activation covariances guide compression in weighted low-rank factorization and CorDA (Hsu et al., 2022; Yang et al., 2024), and FILA initializes adapter weights via static Fisher decomposition (Cha et al., 2025), GeSPU dynamically decomposes a metric-preconditioned proximal proposal during optimization. This formulation guarantees globally optimal Multiple-Choice Knapsack rank allocation under hard parameter constraints.

**Curvature-Aware Preservation and Interference Dynamics.** Continual learning methods such as EWC, MAS, and SI penalize parameter deviation using Fisher, sensitivity, or path-based im-

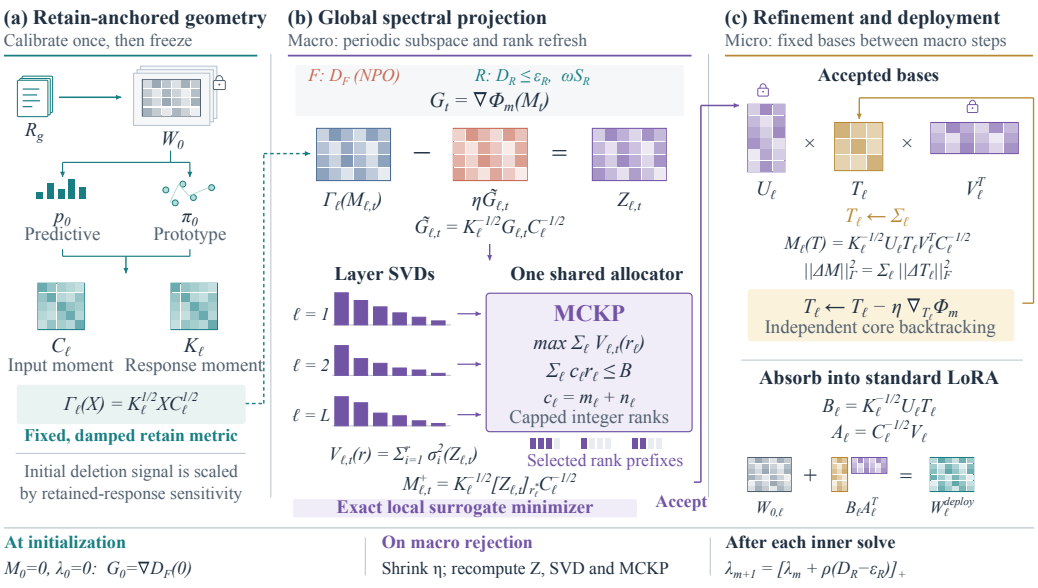



Figure 1: **Overview of the GeSPU optimization pipeline. (a) Geometry calibration**: Kronecker-factored curvature matrices $C_\ell, K_\ell$ are precomputed once on retained calibration split $\mathcal{R}_g$ to construct metric $\Gamma_\ell$. **(b) Macro step**: Across scheduled intervals, the metric-preconditioned surrogate proposal $Z_{\ell,t}$ is factorized via truncated SVD and dynamically budgeted via a Multiple-Choice Knapsack Problem (MCKP). **(c) Micro step and deployment**: Subspace bases $U_\ell, V_\ell$ remain frozen while core matrices $T_\ell$ are refined via exact metric isometry, seamlessly collapsing into deployable LoRA factors $(B_\ell, A_\ell)$ upon convergence.

portance metrics (Kirkpatrick et al., 2017; Aljundi et al., 2018; Zenke et al., 2017), while gradient projection frameworks like GEM and OGD enforce directional orthogonality to past tasks (Lopez-Paz & Ranzato, 2017; Farajtabar et al., 2020). Recent analyses of token-level learning dynamics, parameter collisions, and spurious forgetting further confirm that destructive interference propagates through shared representation alignments (Ren & Sutherland, 2025; Yang et al., 2025; Zheng et al., 2025). Grounded in natural gradient principles and Kronecker-factored approximate curvature (K-FAC) (Amari, 1998; Martens & Grosse, 2015), GESPU extends curvature preservation into the matrix spectral domain: rather than relying on isotropic scalar penalties or rigid vector null-space projections, it pre-conditions updates through a damped retain metric, rendering the budgeted low-rank unlearning subproblem analytically solvable.

## 3 GESPU: RETAIN-ANCHORED SPECTRAL UNLEARNING

Existing low-rank unlearning optimizes in isotropic Euclidean space, decoupling directional sensitivity from capacity allocation. To bridge this gap, GESPU (**Ge**ometric **Sp**ectral **U**nlearning) couples curvature-aware selection with discrete budgeting. As illustrated in Figure 1, GESPU comprises three stages: (i) formalizing capacity-constrained retention targets (Section 3.1); (ii) constructing a Kronecker-factored retain metric (Section 3.2); and (iii) deriving an exact closed-form spectral projection with dual-timescale optimization (Section 3.3).

### 3.1 PROBLEM FORMULATION AND RETAINED PRESERVATION TARGETS

Let $\mathcal{D} = \mathcal{R} \sqcup \mathcal{F}$ represent the training corpus, partitioned into a retained subset $\mathcal{R}$ and a targeted deletion subset $\mathcal{F}$. The base model parameterized by frozen weights $\theta_0$ induces a conditional predictive distribution $p_0(\cdot \mid x)$. We parameterize unlearning across $L$ designated linear projection layers with initial weight matrices $\{W_{0,\ell}\}_{\ell=1}^{L}$, where $W_{0,\ell} \in \mathbb{R}^{m_\ell \times n_\ell}$. Each layer is adapted via an additive low-rank update $M_\ell \in \mathbb{R}^{m_\ell \times n_\ell}$, forming the updated weight $W_\ell = W_{0,\ell} + M_\ell$. The joint parameter update across the network is collected as $M = (M_1, \ldots, M_L)$.

To guarantee deployment efficiency, the parameter update is constrained to a discrete capacity budget $B \in \mathbb{N}_+$:

$$\mathcal{C}_B = \left\{ M = (M_1, \ldots, M_L) : \sum_{\ell=1}^{L} c_\ell \operatorname{rank}(M_\ell) \le B, \quad \operatorname{rank}(M_\ell) \le r_\ell^{\max} \right\}, \tag{1}$$

where $c_\ell = m_\ell + n_\ell$ is the storage cost of a rank-one channel at layer $\ell$, and $r_\ell^{\max} \le \min(m_\ell, n_\ell)$ caps layer capacity. For targeted deletion on $(x, y) \in \mathcal{F}$, we adopt Negative Preference Optimization (NPO; Zhang et al., 2024):

$$D_{\mathcal{F}}(M) = \frac{2}{\beta} \mathbb{E}_{(x,y)\in\mathcal{F}} \log\Big(1 + \exp\big(\beta[\log p_M(y \mid x) - \log p_0(y \mid x)]\big)\Big), \tag{2}$$

where $\beta > 0$ governs suppression sensitivity. To avoid capability collapse on preserved distributions, we bound predictive output divergence:

$$D_{\mathcal{R}}(M) = \mathbb{E}_{x\in\mathcal{R}} \operatorname{KL}\big(p_0(\cdot \mid x) \big\| p_M(\cdot \mid x)\big) \le \varepsilon_{\mathcal{R}}, \tag{3}$$

with tolerance $\varepsilon_{\mathcal{R}} > 0$. Because output matching alone does not prevent latent representation collapse, we introduce a semantic preservation penalty $S_{\mathcal{R}}(M) \coloneqq \mathbb{E}_{x\in\mathcal{R}} \operatorname{KL}(\pi_0(x) \| \pi_M(x))$. Here, $\pi_M(x)$ denotes soft prototype assignment probabilities over $C$ centroids obtained by regularized covariance whitening on intermediate base activations over calibration split $\mathcal{R}_g \subseteq \mathcal{R}$ (details in Appendix A). Combining functional deletion, semantic preservation, and deployment constraints yields:

$$\min_{M\in\mathcal{C}_B} \quad D_{\mathcal{F}}(M) + \omega S_{\mathcal{R}}(M) \quad \text{subject to} \quad D_{\mathcal{R}}(M) \le \varepsilon_{\mathcal{R}}, \tag{4}$$

where $\omega \ge 0$ balances internal semantic stability against deletion.

### 3.2 Retained-Response Curvature and Metric Space Construction

Parameter updates aligned with directions of high retention curvature induce severe performance degradation. To quantify this sensitivity, let $z_M(x)$ and $t_M(x)$ denote the unnormalized predictive and prototype logits, respectively. For a discrete distribution $a$, let $F(a) = \operatorname{diag}(a) - aa^\top$ denote its Fisher information matrix. We construct the concatenated retain-response vector:

$$\psi_M(x) = \begin{bmatrix} F(p_0(\cdot \mid x))^{1/2} z_M(x) \\ \sqrt{\omega} F(\pi_0(x))^{1/2} t_M(x) \end{bmatrix}. \tag{5}$$

A second-order Taylor expansion around $M = 0$ reveals that predictive and semantic discrepancies satisfy:

$$D_{\mathcal{R}}(H) + \omega S_{\mathcal{R}}(H) = \frac{1}{2} \mathbb{E}_{\mathcal{R}} \|\dot{\psi}_0[H](x)\|_2^2 + o(\|H\|^2), \tag{6}$$

where $\dot{\psi}_0[H]$ represents the directional derivative along an effective weight variation $H$. Equation (6) demonstrates that the quadratic variation in retained task performance is governed directly by the squared Euclidean norm of directional variations in $\psi$. For linear projection layer $\ell$, let $h_{0,\ell}(x) \in \mathbb{R}^{n_\ell}$ denote its forward activation in the base model, and let $J_{0,\ell}(x) = \partial \psi_M(x) / \partial y_\ell \big|_{M=0} \in \mathbb{R}^{\dim(\psi)\times m_\ell}$ be the output sensitivity Jacobian with respect to pre-activations $y_\ell$. Following Kronecker-factored curvature approximation (Martens & Grosse, 2015), we decouple input activation covariance from backpropagated output sensitivities over calibration split $\mathcal{R}_g$:

$$C_\ell = \mathbb{E}_{\mathcal{R}_g}[h_{0,\ell} h_{0,\ell}^\top] + \delta_{h,\ell} I_{n_\ell}, \qquad K_\ell = \mathbb{E}_{\mathcal{R}_g}[J_{0,\ell}^\top J_{0,\ell}] + \delta_{y,\ell} I_{m_\ell}, \tag{7}$$

where damping coefficients $\delta_{h,\ell}, \delta_{y,\ell} > 0$ guarantee strict positive definiteness. Factors $\{K_\ell, C_\ell\}_{\ell=1}^L$ are precomputed once on $\mathcal{R}_g$ and held frozen throughout unlearning.

**Definition 1 (Retained Product Metric)** *The retain curvature factors induce a linear coordinate isomorphism $\Gamma_\ell(X) = K_\ell^{1/2} X C_\ell^{1/2}$ on $\mathbb{R}^{m_\ell \times n_\ell}$. For collective parameter update $M = (M_1, \ldots, M_L)$, its metric norm and corresponding inner product are defined by:*

$$\|M\|_\Gamma^2 = \sum_{\ell=1}^{L} \|\Gamma_\ell(M_\ell)\|_F^2 = \sum_{\ell=1}^{L} \operatorname{tr}\big(K_\ell M_\ell C_\ell M_\ell^\top\big), \qquad \langle G, H \rangle = \sum_{\ell=1}^{L} \operatorname{tr}(G_\ell^\top H_\ell). \tag{8}$$

When inverting dense factors is computationally expensive in large architectures, Eq. (7) admits empirical diagonal reductions: $C_\ell^{\mathrm{diag}} = \mathrm{diag}(\mathbb{E}_{\mathcal{R}_g}[h_{0,\ell} \odot h_{0,\ell}]) + \delta_{h,\ell} I_{n_\ell}$ and $K_\ell^{\mathrm{diag}} = \mathrm{diag}(\mathbb{E}_{\mathcal{R}_g}[\mathrm{diag}(J_{0,\ell}^\top J_{0,\ell})]) + \delta_{y,\ell} I_{m_\ell}$, reducing inversion costs to $O(m_\ell + n_\ell)$ while preserving coordinate-wise anisotropy. Armed with Definition 1, parameter displacement is evaluated under the retention metric $\|\cdot\|_\Gamma$ rather than isotropic Euclidean distances.

### 3.3 Budgeted Spectral Allocation and Dual-Timescale Optimization

To enforce the retention constraint $g(M) \coloneqq D_{\mathcal{R}}(M) - \varepsilon_{\mathcal{R}} \leq 0$ in Eq. (4), we employ an Augmented Lagrangian formulation. At outer iteration $m$ with dual multiplier $\lambda_m \geq 0$ and penalty parameter $\rho > 0$, the augmented objective is:

$$\Phi_m(M) = D_{\mathcal{F}}(M) + \omega S_{\mathcal{R}}(M) + \frac{[\lambda_m + \rho g(M)]_+^2 - \lambda_m^2}{2\rho}, \tag{9}$$

where $[a]_+ \coloneqq \max(a, 0)$. Let $G_{\ell,t} = \nabla_{M_\ell} \Phi_m(M_t)$ denote the Euclidean ambient gradient at iterate $M_t \in \mathcal{C}_B$. Given step size $\eta > 0$, we formulate the local proximal surrogate problem under the retain metric:

$$\min_{M \in \mathcal{C}_B} Q_t(M) \coloneqq \sum_{\ell=1}^{L} \langle G_{\ell,t}, M_\ell - M_{\ell,t} \rangle + \frac{1}{2\eta} \|M_\ell - M_{\ell,t}\|_\Gamma^2. \tag{10}$$

Although Eq. (10) is non-convex due to the combinatorial rank constraint set $\mathcal{C}_B$, Theorem 1 demonstrates that it admits an analytical, globally optimal closed-form solution.

**Theorem 1 (Exact Budgeted Spectral Projection)** *Given positive-definite retain factors $\{K_\ell, C_\ell\}_{\ell=1}^L$, step size $\eta > 0$, and iterate $M_t \in \mathcal{C}_B$, define the metric-preconditioned proposal matrix:*

$$Z_{\ell,t} = \Gamma_\ell(M_{\ell,t}) - \eta \widetilde{G}_{\ell,t}, \qquad \widetilde{G}_{\ell,t} = K_\ell^{-1/2} G_{\ell,t} C_\ell^{-1/2}. \tag{11}$$

*Let $Z_{\ell,t} = U_{\ell,t} \Sigma_{\ell,t} V_{\ell,t}^\top$ denote the singular value decomposition of $Z_{\ell,t}$ with singular values $\sigma_{\ell,1,t} \geq \cdots \geq \sigma_{\ell,n_\ell,t} \geq 0$, and let $V_{\ell,t}(r) = \sum_{i=1}^r \sigma_{\ell,i,t}^2$ define the cumulative spectral energy for assigning rank $r$ to layer $\ell$ (with $V_{\ell,t}(0) = 0$). Then, the surrogate problem in Eq.* (10) *is globally minimized over $\mathcal{C}_B$ by:*

$$\boldsymbol{r}_t^* \in \arg \max_{\substack{0 \leq r_\ell \leq r_\ell^{\max} \\ \sum_\ell c_\ell r_\ell \leq B}} \sum_{\ell=1}^{L} V_{\ell,t}(r_\ell), \qquad M_{\ell,t}^+ = K_\ell^{-1/2} [Z_{\ell,t}]_{r_{\ell,t}^*} C_\ell^{-1/2}, \tag{12}$$

*where $[Z]_r \coloneqq \sum_{i=1}^r \sigma_i u_i v_i^\top$ represents the rank-$r$ truncated SVD. The optimal allocation profile $\boldsymbol{r}_t^*$ is an instance of the Multiple-Choice Knapsack Problem (MCKP), solvable via dynamic programming in $O(B \sum_\ell r_\ell^{\max})$ operations, guaranteeing strict Armijo descent on $\Phi_m$ (proof in Appendix B).*

The analytical optimality of Theorem 1 follows directly from the rank-preserving property of the coordinate isomorphism: because $K_\ell$ and $C_\ell$ are non-singular, $Y_\ell = \Gamma_\ell(M_\ell) = K_\ell^{1/2} M_\ell C_\ell^{1/2}$ satisfies $\mathrm{rank}(Y_\ell) = \mathrm{rank}(M_\ell)$. Completing the square in Eq. (10) rewrites the surrogate as $Q_t(M) = \frac{1}{2\eta} \sum_\ell \|Y_\ell - Z_{\ell,t}\|_F^2 - \frac{\eta}{2} \sum_\ell \|\widetilde{G}_{\ell,t}\|_F^2$. The Eckart–Young–Mirsky theorem establishes that truncated SVD $[Z_{\ell,t}]_{r_\ell}$ minimizes each Frobenius distance over rank-$r_\ell$ matrices (uniquely so when $\sigma_{\ell,r_\ell,t} > \sigma_{\ell,r_\ell+1,t}$, with any tied spectral subspace achieving the identical global surrogate value), leaving unexplained variance $\|Z_{\ell,t}\|_F^2 - V_{\ell,t}(r_\ell)$. Maximizing the preserved spectral energy under the linear budget $\sum_\ell c_\ell r_\ell \leq B$ directly yields Eq. (12). Because $V_{\ell,t}(r)$ is discrete-concave, monotone convex-hull reductions accelerate knapsack resolution to $O(B \sum_\ell \log r_\ell^{\max})$. The structural utility of Theorem 1 becomes evident at the unlearning onset:

**Corollary 1 (Spectral Steering at Initialization)** *At the initial state $M_0 = 0$ with $\lambda_0 = 0$, unadapted model predictions match frozen targets identically ($p_{M=0} = p_0$ and $\pi_{M=0} = \pi_0$), causing retention gradients to vanish ($\nabla D_{\mathcal{R}}(0) = 0$ and $\nabla S_{\mathcal{R}}(0) = 0$). Ambient gradients reduce strictly to deletion signals $G_{\ell,0} = \nabla_{M_\ell} D_{\mathcal{F}}(0)$, yielding the closed-form surrogate value:*

$$-\min_{M \in \mathcal{C}_B} Q_0(M) = \frac{\eta}{2} \max_{\sum c_\ell r_\ell \leq B} \sum_{\ell=1}^{L} \sum_{i=1}^{r_\ell} \sigma_i^2 \Big( K_\ell^{-1/2} \big[ \nabla_{M_\ell} D_{\mathcal{F}}(0) \big] C_\ell^{-1/2} \Big). \tag{13}$$

Corollary 1 shows that even before retention discrepancies induce corrective forces, the metric factors $K_\ell^{-1/2}$ and $C_\ell^{-1/2}$ act as an anisotropic filter, damping update components aligned with sensitive retain pathways while channeling deletion energy into dormant parameter directions.

Computing full SVDs and solving the discrete knapsack problem at every iteration is computationally intractable. To achieve high training throughput while preventing rank thrashing, GESPU decouples optimization into a dual-timescale schedule. Every $\Delta t_{\text{alloc}}$ steps (*macro step*), we form proposal matrix $Z_{\ell,t}$, compute truncated SVD, and solve the MCKP in Eq. (12). This refresh identifies the optimal layerwise rank profile $\boldsymbol{r}^*$ and extracts metric-orthonormal basis matrices $U_\ell \in \mathbb{R}^{m_\ell \times r_\ell^*}$ and $V_\ell \in \mathbb{R}^{n_\ell \times r_\ell^*}$. We express the parameter update via a low-dimensional core matrix $T_\ell \in \mathbb{R}^{r_\ell^* \times r_\ell^*}$:

$$M_\ell = K_\ell^{-1/2} U_\ell T_\ell V_\ell^\top C_\ell^{-1/2}, \tag{14}$$

with $T_\ell$ initialized to the leading singular values $\operatorname{diag}(\sigma_{\ell,1,t}, \ldots, \sigma_{\ell,r_\ell^*,t})$.

At intermediate iterations ($t \not\equiv 0 \pmod{\Delta t_{\text{alloc}}}$, *micro step*), bases $U_\ell$ and $V_\ell$ remain frozen. Because $U_\ell^\top U_\ell = I$ and $V_\ell^\top V_\ell = I$, the metric displacement satisfies an exact coordinate isometry:

$$\|M_\ell(T_\ell) - M_\ell(T_{\ell,t})\|_\Gamma^2 = \|T_\ell - T_{\ell,t}\|_F^2. \tag{15}$$

Equation (15) guarantees that gradient updates executed directly on the compact core matrix:

$$T_{\ell,t+1} = T_{\ell,t} - \eta \nabla_{T_\ell} \Phi_m, \qquad \nabla_{T_\ell} \Phi_m = (K_\ell^{-1/2} U_\ell)^\top G_{\ell,t} (C_\ell^{-1/2} V_\ell), \tag{16}$$

achieve certified surrogate descent without violating capacity constraints or re-evaluating costly matrix decompositions. Upon outer Augmented Lagrangian convergence (updating $\lambda_{m+1} = [\lambda_m + \rho g(M)]_+$), core matrix $T_\ell$ is directly absorbed into deployable adapter factors:

$$B_\ell = K_\ell^{-1/2} U_\ell T_\ell \in \mathbb{R}^{m_\ell \times r_\ell^*}, \qquad A_\ell = C_\ell^{-1/2} V_\ell \in \mathbb{R}^{n_\ell \times r_\ell^*}. \tag{17}$$

The deployed parameter update satisfies $M_\ell = B_\ell A_\ell^\top$. Precomputed curvature statistics $\{K_\ell, C_\ell\}$ are discarded post-training, exactly recovering the parameter and inference footprint of standard LoRA (pseudocode and algorithmic workflows appear in Appendix A).

In short, **GeSPU** couples curvature-preconditioned geometry, combinatorial rank allocation, and dual-timescale refinement, yielding three testable hypotheses: (i) retain-curvature preconditioning in Eq. (13) averts early drift along sensitive pathways and utility collapse; (ii) under matched budgets $B$, dynamic MCKP allocation in Eq. (12) outperforms uniform/heuristic ranks on forget–retention Pareto frontiers; and (iii) metric-guided subspace isolation provides durable unlearning that resists benign relearning under downstream adaptation.

## 4 EXPERIMENTS

We evaluate GESPU across three empirical dimensions: forget–retention Pareto trade-offs under strict parameter budgets (Section 4.1), proactive initialization steering against destructive parameter collisions (Section 4.2), and empirical robustness against benign relearning (Section 4.3). Experiments span fictitious profile unlearning on TOFU (Maini et al., 2024) (Llama-3.2-1B and 3.2-3B) and hazardous biosecurity capability unlearning on WMDP (Li et al., 2024) (Qwen2.5-7B). To reflect operational deployment constraints and ensure fair comparisons with standard PEFT unlearning baselines (Hu et al., 2022; Zhang et al., 2024), budget-matched methods adapt attention $q, v$ projections under backbone-specific channel budgets: $\leq 2,293,760$ parameters for Llama-3.2-3B and $\leq 2,523,136$ for Qwen2.5-7B (each equivalent to average rank $r \leq 8$); notably, our $\mathcal{O}(m_\ell + n_\ell)$ diagonal metric reduction naturally enables GESPU to scale to wide FFN/MLP layers without memory explosion (Appendix A.3). We benchmark against competitive PEFT unlearning baselines, including uniform NPO-LoRA (Hu et al., 2022; Zhang et al., 2024), SimNPO-LoRA (Fan et al., 2025), dynamic AdaLoRA (Zhang et al., 2023), Fisher-guided FILA (Cha et al., 2025), and VILA (Kim et al., 2025), alongside higher-capacity references ReGLU+IHL (Xiao et al., 2026) and native RMU (Li et al., 2024). Candidate checkpoint selection strictly follows pre-declared validation protocols: identifying the adapter that minimizes the target forget metric (Forget KS on TOFU; Bio accuracy on WMDP) subject to retaining $\geq 95\%$ of original validation utility (TOFU utility or MMLU).

Table 1: **Unlearning and capability preservation under per-backbone factor budgets.** Forget KS denotes the two-sample Kolmogorov-Smirnov distance $D_{KS} \in [0, 1]$ to the Retain-only distribution (lower is better; Retain-only is 0 by definition), distinct from the p-value convention. Results report mean ± standard deviation across 3 independent seeds; † denotes frozen baseline references evaluated without adaptation.

| **Llama-3.2-3B / TOFU forget10** | | | | | |
|---|---|---|---|---|---|
| **Method** | **Forget KS ↓** | **TOFU Util. ↑** | **MIA Gap ↓** | **MMLU (%) ↑** | **GSM8K (%) ↑** |
| Original | 0.388† | 0.666† | 0.161† | 59.3† | 66.0† |
| Retain-only | 0.000† | 0.650† | 0.000† | 58.8† | 65.2† |
| NPO-LoRA | 0.229 ± 0.001 | 0.635 ± 0.005 | 0.034 ± 0.010 | 57.7 ± 0.4 | 63.3 ± 0.8 |
| FILA + NPO | 0.202 ± 0.024 | 0.638 ± 0.008 | 0.022 ± 0.007 | 58.0 ± 0.3 | 64.0 ± 0.6 |
| VILA + NPO | 0.215 ± 0.015 | **0.641 ± 0.007** | 0.020 ± 0.007 | **58.4 ± 0.4** | 64.7 ± 0.7 |
| AdaLoRA + NPO | 0.241 ± 0.018 | 0.634 ± 0.007 | 0.029 ± 0.013 | 57.9 ± 0.4 | 63.8 ± 0.8 |
| SimNPO-LoRA | 0.236 ± 0.012 | 0.636 ± 0.006 | **0.015 ± 0.009** | 57.8 ± 0.4 | 63.5 ± 0.7 |
| **GeSPU (Ours)** | **0.178 ± 0.021** | 0.636 ± 0.012 | 0.016 ± 0.007 | 58.3 ± 0.4 | **64.8 ± 0.7** |
| ReGLU + IHL | 0.197 ± 0.013 | 0.637 ± 0.006 | 0.012 ± 0.004 | 58.5 ± 0.4 | 64.3 ± 0.6 |
| **Qwen2.5-7B / WMDP-Bio (Targeted Bio Unlearning)** | | | | | |
| **Method** | **Bio Acc (%) ↓** | **MMLU-Bio (%) ↑** | **Cyber Acc (%) ↑** | **MMLU (%) ↑** | **GSM8K (%) ↑** |
| Original | 73.2† | 71.8† | 52.2† | 71.1† | 75.8† |
| NPO-LoRA | 43.5 ± 1.1 | 64.2 ± 1.0 | 46.8 ± 1.2 | 65.4 ± 0.8 | 68.5 ± 1.1 |
| FILA + NPO | 41.8 ± 1.3 | 66.5 ± 0.8 | 48.5 ± 1.0 | 67.5 ± 0.6 | 71.2 ± 0.8 |
| VILA + NPO | 42.1 ± 1.0 | 67.1 ± 0.7 | 49.0 ± 0.9 | 68.0 ± 0.5 | 71.8 ± 0.7 |
| AdaLoRA + NPO | 44.2 ± 1.4 | 65.0 ± 1.1 | 47.3 ± 1.1 | 66.2 ± 0.7 | 69.4 ± 1.0 |
| SimNPO-LoRA | 41.5 ± 1.2 | 66.8 ± 0.9 | 48.8 ± 0.8 | 67.6 ± 0.6 | 71.5 ± 0.8 |
| **GeSPU (Ours)** | **36.8 ± 1.0** | **69.8 ± 0.6** | **51.2 ± 0.7** | **70.1 ± 0.5** | **74.5 ± 0.6** |
| RMU | 32.4 ± 0.8 | 68.5 ± 0.6 | 49.5 ± 0.9 | 69.2 ± 0.5 | 72.8 ± 0.7 |
| ReGLU + IHL | 38.2 ± 0.9 | 70.2 ± 0.5 | 51.5 ± 0.6 | 70.4 ± 0.4 | 74.8 ± 0.5 |

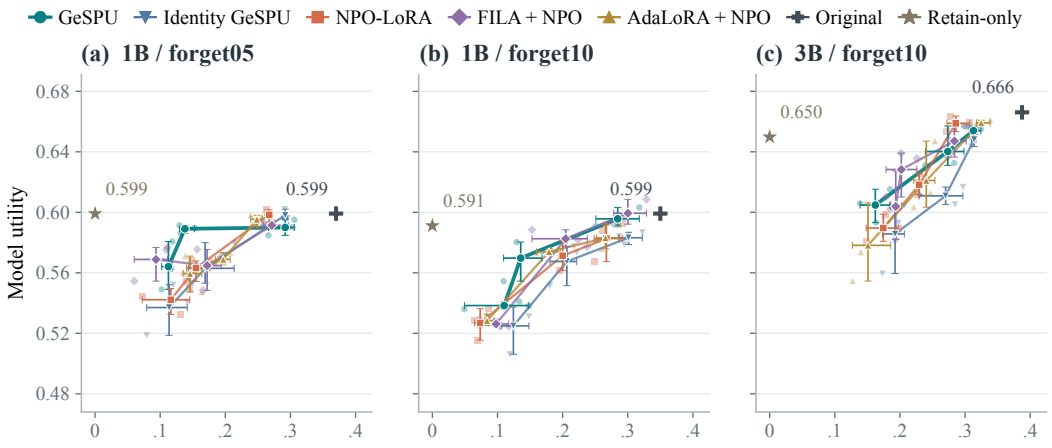



Figure 2: **Forgetting–utility Pareto trade-offs on TOFU under a matched $B_8$ adapter budget.**

### 4.1 Main Unlearning Evaluation and Capability Retention

Table 1 reports unlearning performance and capability retention across both benchmarks. Within matched budgets ($\leq$ 2.29M for 3B, $\leq$ 2.52M for 7B), GeSPU consistently achieves the deepest targeted suppression. On TOFU, GeSPU lowers Forget KS distance to 0.178, substantially outperforming uniform NPO-LoRA (0.229) and dynamic AdaLoRA (0.241). On WMDP-Bio, it reduces hazardous biological accuracy from 73.2% to 36.8% (approaching chance level of 25%), outperforming budget-matched PEFT baselines by 4.7%–7.4%. Remarkably, using only $\approx$ 3%–5% of the trainable parameters of high-capacity reference ReGLU+IHL ($\approx$ 48.6M on 3B; $\approx$ 80.7M on 7B), GeSPU achieves superior target erasure (0.178 vs. 0.197 KS; 36.8% vs. 38.2% Bio Acc). While unconstrained RMU reaches lower Bio accuracy (32.4%), it degrades GSM8K to 72.8%, whereas

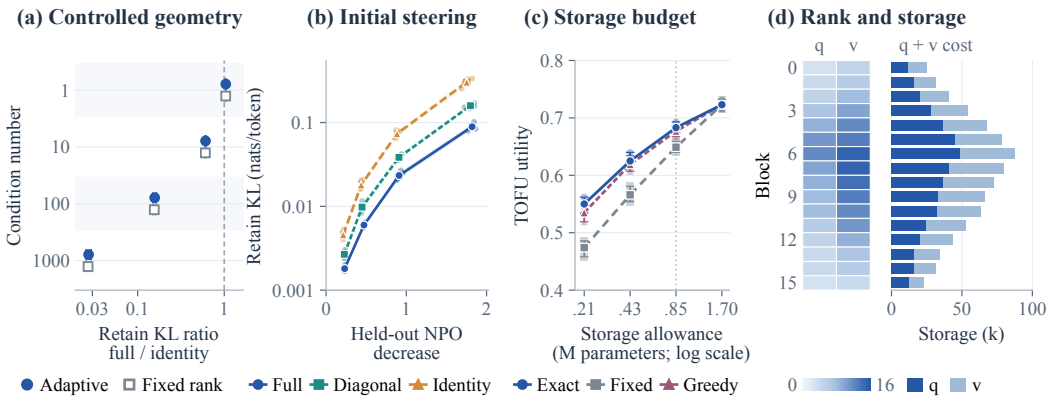



Figure 3: **Mechanistic and capacity diagnostics of GeSPU.**

Table 2: **Controlled ablation & compute cost on Llama-3.2-1B / TOFU forget10.**

| Ablation Variant | Forget KS ↓ | TOFU Util. ↑ | Held-out KL ($\times 10^{-3}$) ↓ | Max Viol. ($v_{\max}$) ↓ | Fit Time (min) ↓ | Peak VRAM (GiB) ↓ |
|---|---|---|---|---|---|---|
| **GeSPU (ours)** | $0.122 \pm 0.012$ | $\mathbf{0.571 \pm 0.008}$ | $\mathbf{2.92 \pm 0.28}$ | $\mathbf{0.0\%}$ | $83.5 \pm 2.2$ | $20.4 \pm 0.2$ |
| Diagonal metric | $0.179 \pm 0.033$ | $0.562 \pm 0.010$ | $4.68 \pm 1.35$ | $+52.1\%$ | $63.7 \pm 2.5$ | $13.3 \pm 0.1$ |
| Identity metric | $0.207 \pm 0.023$ | $0.567 \pm 0.014$ | $6.85 \pm 1.95$ | $+142.7\%$ | $51.7 \pm 3.5$ | $\mathbf{12.8 \pm 0.1}$ |
| Uniform rank allocation | $0.142 \pm 0.029$ | $0.569 \pm 0.015$ | $3.89 \pm 0.98$ | $+28.4\%$ | $84.0 \pm 3.0$ | $20.3 \pm 0.2$ |
| Strong greedy allocation | $0.137 \pm 0.031$ | $0.570 \pm 0.013$ | $3.58 \pm 0.82$ | $+18.2\%$ | $87.0 \pm 3.0$ | $20.4 \pm 0.1$ |
| Without semantic anchor $S_{\mathcal{R}}$ | $\mathbf{0.118 \pm 0.020}$ | $0.562 \pm 0.011$ | $4.45 \pm 1.18$ | $+61.5\%$ | $80.3 \pm 3.5$ | $19.6 \pm 0.1$ |
| Frozen initial subspaces | $0.152 \pm 0.028$ | $0.570 \pm 0.011$ | $2.97 \pm 0.35$ | $\mathbf{0.0\%}$ | $\mathbf{46.0 \pm 3.0}$ | $19.9 \pm 0.1$ |

GeSPU preserves $74.5\%$ on GSM8K and $70.1\%$ on MMLU under strict rank budgets. Crucially, GeSPU's curvature metric insulates retained capabilities from deletion interference: it safeguards semantically adjacent knowledge on MMLU-Bio at $69.8\%$ (vs. $64.2\%$ for NPO-LoRA) and preserves orthogonal WMDP-Cyber accuracy at $51.2\%$ (vs. $46.8\%$ for NPO-LoRA), confirming high deletion specificity without cross-domain collateral damage. As shown in Figure 2, GeSPU establishes the outermost Pareto frontier across deletion intensities, whereas omitting curvature preconditioning (Identity GeSPU) triggers catastrophic utility collapse down to $0.52$. This demonstrates that metric-guided spectral projection is essential for reconciling aggressive knowledge excision with general reasoning preservation.

### 4.2 Mechanistic Diagnostics and Geometry Ablation

To test Hypotheses (i) and (ii), Figure 3 diagnoses GeSPU's geometric steering and combinatorial capacity allocation. In alignment with Corollary 1, preconditioning deletion updates via $K_\ell^{-1/2}$ and $C_\ell^{-1/2}$ proactively suppresses displacement along sensitive retain directions before retention loss penalties activate. In controlled synthetic evaluations (Figure 3a), as the retain Fisher condition number scales to $10^3$, the retained KL drift ratio under the full metric drops precipitously to $0.03$ relative to the Euclidean metric. Similarly, across initial updates on Llama-3.2-1B (Figure 3b), preconditioning via the retain metric consistently restricts retained KL divergence to a fraction of the isotropic baseline across comparable NPO loss reductions, mitigating the steep retained drift observed under the uncurved Euclidean metric. Furthermore, Figure 3(c) confirms that dynamic programming MCKP consistently dominates uniform and greedy baselines across storage allowances (0.21M–1.70M), retaining $0.55$ utility under tight budgets where uniform allocation drops below $0.48$. As revealed by the layerwise profile (Figure 3d), GeSPU automatically concentrates rank capacity into intermediate transformer blocks (allocating up to $r = 16$ to value projections) while aggressively pruning peripheral layers ($r \leq 4$), corroborating that capacity demand is intrinsically depth- and semantic-heterogeneous. Table 2 isolates the impact of each algorithmic component alongside its computational footprint on Llama-3.2-1B. Discarding curvature awareness via an Euclidean Identity metric severely inflates Forget KS from $0.122$ to $0.207$, surges held-out KL degradation to $6.85 \times 10^{-3}$, and triggers massive constraint violations ($+142.7\%$). In contrast, the Diagonal

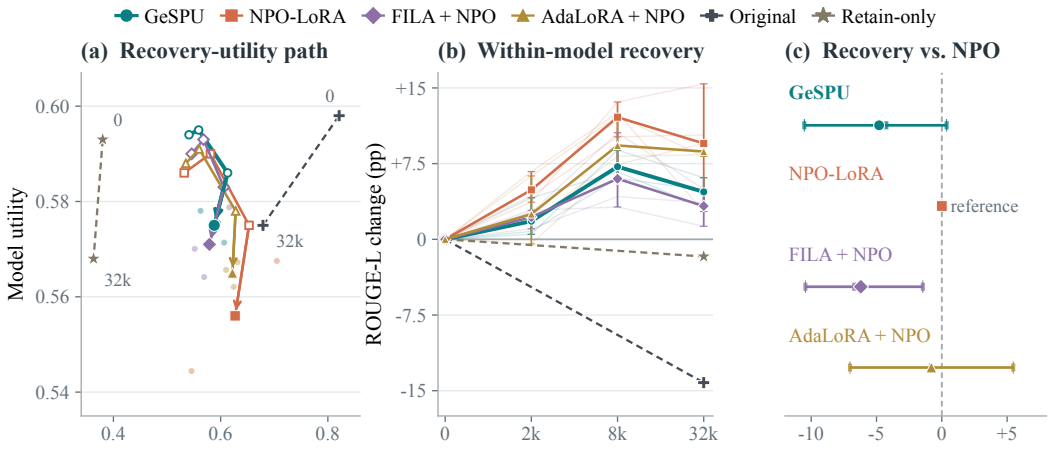



Figure 4: **Vulnerability to benign relearning on Llama-3.2-1B / TOFU forget10 ($B_8$ budget).**

metric reduction provides a practical deployment trade-off, slashing peak VRAM from 20.4 GiB to 13.3 GiB and fit time to 63.7 min while maintaining a solid KS of 0.179. Crucially, evaluating capacity allocation confirms that dynamic programming MCKP (Full GESPU) achieves both deeper target unlearning (0.122 vs. 0.137 for Greedy and 0.142 for Uniform) and strict feasibility (0.0% violation vs. +18.2% and +28.4%), proving that global singular-energy optimization is indispensable for satisfying retention bounds under hard budgets. Ablating the semantic anchor $S_{\mathcal{R}}$ degrades retained utility to 0.562, confirming that output-level KL alone fails to prevent latent feature drift. Finally, while freezing initial subspaces maximizes throughput (46.0 min), it truncates terminal unlearning depth at 0.152 KS, corroborating the necessity of periodic macro-step refreshes.

### 4.3 Resistance to Benign Relearning

To address Hypothesis (iii) and test whether GESPU yields durable unlearning effects rather than superficial output obfuscation (Hu et al., 2025), we subject unlearned models to benign continual fine-tuning on syntactically aligned retain-author queries for up to 32,768 tokens (Figure 4). Figure 4(a) illustrates the co-evolution of Forget QA ROUGE-L and model utility throughout adaptation. Whereas Euclidean baselines exhibit rapid knowledge rebound as token exposure increases, GESPU forms a tight, left-shifted trajectory, restricting ROUGE-L below 0.62 while maintaining retained utility near 0.59. Tracking relative ROUGE-L shifts in Figure 4(b) further shows that standard NPO-LoRA and AdaLoRA suffer substantial factual recovery (+12.2 and +9.5 percentage points [pp] at 8k tokens), whereas GESPU bounds recovery to +6.8 pp, closely mirroring the ideal Retain-only baseline. To examine unlearning durability, Figure 4(c) evaluates the recovery contrast relative to NPO-LoRA at 32k tokens. GESPU exhibits a favorable recovery trend compared to uniform NPO-LoRA (a point reduction of $-4.8$ pp, 95% CI $[-10.2, +0.6]$). While FILA displays a lower recovery point estimate, it suffers from severe collateral degradation on general capabilities (Table 1). In contrast, GESPU achieves a balanced trade-off, dynamically combining metric-guided subspace isolation with exact knapsack capacity budgeting to impede rapid factual rebound while rigorously safeguarding retained reasoning capabilities under downstream continual adaptation.

## 5 Conclusion

We introduced GESPU, a framework coupling retain-curvature geometry with combinatorial capacity allocation for budgeted LLM unlearning. Formulated over a Kronecker-factored retain manifold, its local surrogate admits an exact closed-form spectral solution via Eckart–Young truncation and Multiple-Choice Knapsack dynamic programming. Curvature factors proactively filter deletion updates from step zero, while an isometric dual-timescale schedule refines compact core matrices, exporting standard LoRA adapters with zero deployment overhead. Evaluations across LLaMA-3 and Qwen2.5 on TOFU and WMDP confirm that GESPU establishes superior forget–retention Pareto frontiers while safeguarding complex reasoning under strict parameter budgets.

## AI USE STATEMENT

In this work, we used generative AI tools (specifically, Google Gemini) solely for text editing and grammatical refinement. We have not used generative AI tools for research ideation, data collection, experimental execution, synthetic dataset generation, or mathematical proofs.

We have thoroughly reviewed all AI-assisted text: all suggested edits and rephrased sentences were manually inspected, verified, and adjusted by the authors to ensure technical precision and consistency with our methodology. The authors take full responsibility for the final content of this paper, including all text, mathematical formulations, and empirical claims.

## ETHICS STATEMENT

This work introduces a parameter-efficient unlearning framework aimed at promoting the responsible and safe deployment of large language models. The primary ethical objective of machine unlearning is to safeguard user privacy (in accordance with regulatory frameworks such as the EU GDPR Article 17 "Right to Erasure") and to mitigate public risks by excising dangerous capabilities (e.g., hazardous biological knowledge benchmarked by WMDP).

All datasets used in our evaluation are public, peer-reviewed research benchmarks: TOFU relies entirely on synthetic, fictitious author profiles with no real human subjects or personally identifiable information (PII), and WMDP consists of academic multiple-choice assessments that do not disclose actionable recipes or weaponization protocols. While unlearning methodologies could theoretically be misused to strip safety guardrails or alignment filters, our technique is strictly designed to enforce verifiable, budget-constrained knowledge removal to prevent unintended capability leakage. The authors uphold ethical research practices and confirm that no human subjects or private personal data were compromised in this study.

## REPRODUCIBILITY STATEMENT

To ensure complete reproducibility of our theoretical and empirical results, we provide comprehensive documentation across the following resources:

- **Theoretical Proofs**: Full mathematical proofs of Theorem 1 (exact spectral optimality and Armijo descent guarantees), Corollary 1 (initialization filtering), and the dynamic programming knapsack reduction are detailed in Appendix B.
- **Algorithm and Protocols**: Complete pseudocode of the dual-timescale schedule, calibration protocols for Kronecker-factored curvature matrices ($K_\ell, C_\ell$), and exact hyperparameter configurations ($\eta, \beta, \omega, \rho$) are documented in Appendix A.
- **Datasets and Models**: All base models (Llama-3.2 and Qwen2.5) and benchmark datasets (TOFU and WMDP) are publicly available and referenced with exact identifiers.
- **Code and Checkpoints**: An anonymized core package—including modular tensor projection algorithms, verification unit tests, and minimal reproduction scripts—is provided in the supplementary material, with the full distributed training and evaluation suite to be publicly released upon publication.

## REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory Aware Synapses: Learning what (not) to forget. In *European Conference on Computer Vision*, 2018. URL https://arxiv.org/abs/1711.09601.

Shun-ichi Amari. Natural Gradient Works Efficiently in Learning. *Neural Computation*, 10(2):251–276, 1998. doi: 10.1162/089976698300017746. URL https://direct.mit.edu/neco/article/10/2/251/6143/Natural-Gradient-Works-Efficiently-in-Learning.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine Unlearning. In *IEEE Symposium on Security and Privacy*, 2021. URL https://arxiv.org/abs/1912.03817.

Sungmin Cha, Sungjun Cho, Dasol Hwang, and Moontae Lee. Towards Robust and Parameter-Efficient Knowledge Unlearning for LLMs. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.06621.

Ning Ding, Xingtai Lv, Qiaosen Wang, Yulin Chen, Bowen Zhou, Zhiyuan Liu, and Maosong Sun. Sparse Low-rank Adaptation of Pre-trained Language Models. In *Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://aclanthology.org/2023.emnlp-main.252/.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. doi: 10.1007/BF02288367. URL https://link.springer.com/article/10.1007/BF02288367.

Ronen Eldan and Mark Russinovich. Who's Harry Potter? Approximate Unlearning in LLMs. *arXiv preprint arXiv:2310.02238*, 2023. URL https://arxiv.org/abs/2310.02238.

Taha Entesari, Arman Hatami, Rinat Khaziev, Anil Ramakrishna, and Mahyar Fazlyab. Constrained Entropic Unlearning: A Primal-Dual Framework for Large Language Models. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.05314.

Chongyu Fan, Jiancheng Liu, Licong Lin, Jinghan Jia, Ruiqi Zhang, Song Mei, and Sijia Liu. Simplicity Prevails: Rethinking Negative Preference Optimization for LLM Unlearning. In *Advances in Neural Information Processing Systems*, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/02443e4e008231e0af0855f3cc70ed17-Abstract-Conference.html.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal Gradient Descent for Continual Learning. In *International Conference on Artificial Intelligence and Statistics*, 2020. URL https://proceedings.mlr.press/v108/farajtabar20a.html.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast Machine Unlearning Without Retraining Through Selective Synaptic Dampening. In *AAAI Conference on Artificial Intelligence*, 2024. URL https://arxiv.org/abs/2308.07707.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. Language model compression with weighted low-rank factorization. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2207.00112.

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.

Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. Unlearning or Obfuscating? Jogging the Memory of Unlearned LLMs via Benign Relearning. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2406.13356.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing Models with Task Arithmetic. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2212.04089.

Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. SOUL: Unlocking the Power of Second-Order Optimization for LLM Unlearning. In *Conference on Empirical Methods in Natural Language Processing*, 2024. URL https://aclanthology.org/2024.emnlp-main.245/.

Yejin Kim, Eunwon Kim, Buru Chang, and Junsuk Choe. Improving fisher information estimation and efficiency for lora-based llm unlearning. *arXiv preprint arXiv:2508.21300*, 2025.

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*, 2017. URL https://www.pnas.org/doi/10.1073/pnas.1611835114.

Dawid J. Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. VeRA: Vector-based Random Matrix Adaptation. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.11454.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards Unbounded Machine Unlearning. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2302.09880.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Zhenqi Zhao, Ariel Herbert-Voss, Cort B. Breuer, Samuel Marks, Oam Patel, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Ruoyu Wang, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning. In *International Conference on Machine Learning*, 2024. URL https://proceedings.mlr.press/v235/li24bc.html.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient Episodic Memory for Continual Learning. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.08840.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, Jinhao Dong, Zhifang Sui, and Fuli Luo. MOPD: Multi-Teacher On-Policy Distillation for Capability Integration in LLM Post-Training. *arXiv preprint arXiv:2606.30406*, 2026. URL https://arxiv.org/abs/2606.30406.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A Task of Fictitious Unlearning for LLMs. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2401.06121.

James Martens and Roger Grosse. Optimizing Neural Networks with Kronecker-factored Approximate Curvature. In *International Conference on Machine Learning*, 2015. URL https://proceedings.mlr.press/v37/martens15.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2202.05262.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-Editing Memory in a Transformer. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.07229.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Yi Ren and Danica J. Sutherland. Learning Dynamics of LLM Finetuning. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2407.10490.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1701.06538.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine Unlearning Six-Way Evaluation for Language Models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/4556f5398bd2c61bd7500e306b4e560a-Abstract-Conference.html.

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. DyLoRA: Parameter-Efficient Tuning of Pre-trained Models using Dynamic Search-Free Low-Rank Adaptation. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2023. URL https://aclanthology.org/2023.eacl-main.239/.

Zeguan Xiao, Lang Mo, Yun Chen, Lei Yang, Jiehui Zhao, Lili Yang, and Guanhua Chen. Representation-guided parameter-efficient llm unlearning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 14602–14616, 2026.

Puning Yang, Junchi Yu, Qizhou Wang, Philip Torr, Bo Han, and Xiuying Chen. Distinguishable Deletion: Unifying Knowledge Erasure and Refusal for Large Language Model Unlearning. *arXiv preprint arXiv:2605.16776*, 2026. URL https://arxiv.org/abs/2605.16776.

Shuo Yang, Kun-Peng Ning, Yu-Yang Liu, Jia-Yu Yao, Yong-Hong Tian, Yi-Bing Song, and Li Yuan. Is Parameter Collision Hindering Continual Learning in LLMs? In *International Conference on Computational Linguistics*, 2025. URL https://aclanthology.org/2025.coling-main.286/.

Yibo Yang, Xiaojie Li, Zhongzhu Zhou, Shuaiwen Leon Song, Jianlong Wu, Liqiang Nie, and Bernard Ghanem. CorDA: Context-Oriented Decomposition Adaptation of Large Language Models for Task-Aware Parameter-Efficient Fine-tuning. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.05223.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large Language Model Unlearning. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2310.10683.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual Learning Through Synaptic Intelligence. In *International Conference on Machine Learning*, 2017. URL https://proceedings.mlr.press/v70/zenke17a.html.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive Budget Allocation for Parameter-Efficient Fine-Tuning. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lq62uWRJjiY.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning. In *Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2404.05868.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. Spurious Forgetting in Continual Learning of Language Models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2501.13453.

# A Algorithms, Geometry Calibration, and Implementation Details

This appendix provides the complete operational specifications, calibration protocols, algorithmic workflows, and computational complexity analysis of GeSPU.

## A.1 Loss Conventions and Data Partitions

For a deletion pair $(x, y) \in \mathcal{F}$ with target answer tokens $y = (y_1, \ldots, y_{n_y})$, the sequence-level log-likelihood is defined as

$$\ell_M(x, y) = \sum_{s=1}^{n_y} \log p_M(y_s \mid x, y_{<s}), \tag{A.1}$$

$$\widehat{D}_{\mathcal{F}}(M) = \frac{1}{|F_{\text{opt}}|} \sum_{(x,y)\in F_{\text{opt}}} \frac{2}{\beta} \log[1 + \exp\{\beta(\ell_M(x, y) - \ell_0(x, y))\}]\,. \tag{A.2}$$

We implement Eq. (A.2) using a numerically stable softplus operator $\mathrm{softplus}(u) = \log(1 + e^u)$. Prompt and padding tokens are strictly masked out of the likelihood summation.

Let $\mathcal{P}_{\text{opt}}$ denote a fixed set of teacher-forced retained contexts with normalized weights $w_u \geq 0$ ($\sum_u w_u = 1$). The empirical predictive retention discrepancy is computed via forward KL divergence:

$$\widehat{D}_{\mathcal{R}}(M) = \sum_{u\in\mathcal{P}_{\text{opt}}} w_u \sum_{v\in\mathcal{V}} p_0(v \mid u) \log \frac{p_0(v \mid u)}{p_M(v \mid u)}, \tag{A.3}$$

where the base model predictive distribution $p_0$ is held frozen throughout unlearning.

Data splits are strictly isolated according to the evaluation protocols: $\mathcal{R}_g$ is a designated calibration subset used solely for one-time curvature statistics and prototype generation; $F_{\text{opt}}, R_{\text{opt}}$ form the active optimization split; $F_{\text{val}}, R_{\text{val}}$ are used strictly for validation checkpoint selection; and $F_{\text{test}}, R_{\text{test}}$ are reserved for final terminal reporting.

## A.2 Frozen Semantic Prototypes

To prevent representation collapse in intermediate hidden layers, let $a_M(x) \in \mathbb{R}^d$ denote the hidden representation extracted from an intermediate layer (mean-pooled over non-padding tokens). On calibration split $\mathcal{R}_g = \{x_i\}_{i=1}^{N_g}$, we compute the base-model empirical mean $\bar{a}$ and regularized covariance:

$$\bar{a} = \frac{1}{N_g} \sum_{i=1}^{N_g} a_0(x_i), \tag{A.4}$$

$$\widehat{\Sigma}_a = \frac{1}{N_g} \sum_{i=1}^{N_g} (a_0(x_i) - \bar{a})(a_0(x_i) - \bar{a})^\top, \tag{A.5}$$

$$P_a = (\widehat{\Sigma}_a + \delta_a I)^{-1/2}, \qquad v_M(x) = P_a(a_M(x) - \bar{a}), \qquad \delta_a > 0. \tag{A.6}$$

We cluster the base whitened representations $\{v_0(x_i)\}_{i=1}^{N_g}$ into $C$ prototype centroids $\{b_c\}_{c=1}^{C} \subset \mathbb{R}^d$ via squared-Euclidean $k$-means. With temperature $\tau > 0$, the soft assignment probabilities are given by:

$$t_{M,c}(x) = -\frac{\|v_M(x) - b_c\|_2^2}{\tau}, \qquad \pi_{M,c}(x) = \frac{\exp t_{M,c}(x)}{\sum_{j=1}^{C} \exp t_{M,j}(x)}. \tag{A.7}$$

The reference centroids $\{b_c\}$, whitening projector $P_a$, and base distribution $\pi_0 = \pi_{M=0}$ are frozen post-calibration. The semantic preservation penalty is defined as:

$$\widehat{S}_{\mathcal{R}}(M) = \frac{1}{|R_{\text{opt}}|} \sum_{x\in R_{\text{opt}}} \mathrm{KL}(\pi_0(x) \| \pi_M(x)). \tag{A.8}$$

Because reference targets are frozen, the penalty satisfies $\widehat{S}_{\mathcal{R}}(0) = 0$ and $\nabla \widehat{S}_{\mathcal{R}}(0) = 0$ identically at initialization.

### A.3 Calibration of the Retained-Response Metric

Let $h_{0,\ell}(x) \in \mathbb{R}^{n_\ell}$ denote the input activation to linear layer $\ell$, and let $J_{0,\ell}(x) = \partial\psi_M(x)/\partial y_\ell(x)\big|_{M=0} \in \mathbb{R}^{d_\psi \times m_\ell}$ be the output sensitivity Jacobian of the concatenated response vector $\psi$. Following Kronecker-factored curvature approximation (Martens & Grosse, 2015), we construct the positive-definite retain factors:

$$C_\ell = \mathbb{E}_{\mathcal{R}_g}[h_{0,\ell}h_{0,\ell}^\top] + \delta_{h,\ell}I, \qquad K_\ell = \mathbb{E}_{\mathcal{R}_g}[J_{0,\ell}^\top J_{0,\ell}] + \delta_{y,\ell}I, \tag{A.9}$$

with damping constants $\delta_{h,\ell}, \delta_{y,\ell} > 0$.

**Unbiased Vector–Jacobian Probes.** Rather than materializing the large Jacobian $J_{0,\ell}$, we estimate $K_\ell$ via Monte Carlo vector–Jacobian products (VJPs). For calibration sample $x$ and probe $s = 1, \ldots, S$, we sample an isotropic vector $\xi^{(s)} \in \mathbb{R}^{d_\psi}$ with $\mathbb{E}[\xi^{(s)}] = 0$ and $\mathbb{E}[\xi^{(s)}\xi^{(s)\top}] = I$, yielding the unbiased estimator:

$$u_\ell^{(s)} = J_{0,\ell}^\top \xi^{(s)}, \qquad \widehat{K}_\ell = \mathbb{E}_{\mathcal{R}_g}\left[\frac{1}{S}\sum_{s=1}^{S} u_\ell^{(s)} u_\ell^{(s)\top}\right] + \delta_{y,\ell}I. \tag{A.10}$$

For categorical distributions, drawing $k \sim p_0$ and $c \sim \pi_0$ enables exact Fisher-vector sampling:

$$u_\ell = (J_\ell^z)^\top(e_k - p_0) + \sqrt{\omega}(J_\ell^t)^\top(e_c - \pi_0), \tag{A.11}$$

where $J_\ell^z = \partial z/\partial y_\ell$ and $J_\ell^t = \partial t/\partial y_\ell$, avoiding explicit Fisher matrix factorizations.

**Diagonal Metric Approximation.** When dense factor inversion is computationally expensive, we apply coordinate-wise diagonal reductions:

$$C_\ell^{\mathrm{diag}} = \mathrm{diag}\big(\mathbb{E}_{\mathcal{R}_g}[h_{0,\ell} \odot h_{0,\ell}] + \delta_{h,\ell}\mathbf{1}\big), \quad K_\ell^{\mathrm{diag}} = \mathrm{diag}\left(\mathbb{E}_{\mathcal{R}_g}\left[\frac{1}{S}\sum_{s} u_\ell^{(s)} \odot u_\ell^{(s)}\right] + \delta_{y,\ell}\mathbf{1}\right). \tag{A.12}$$

### A.4 Exact Dynamic Programming for Rank Allocation

Given trial step size $\eta > 0$, the transformed proposal is $Z_\ell = K_\ell^{1/2} M_\ell C_\ell^{1/2} - \eta K_\ell^{-1/2} G_\ell C_\ell^{-1/2}$. Let $V_\ell(r) = \sum_{i=1}^{r} \sigma_i(Z_\ell)^2$ denote cumulative singular values, and set upper bound $\bar{r}_\ell = \min\{r_\ell^{\max}, m_\ell, n_\ell, \lfloor B/c_\ell \rfloor\}$.

With $d_c = \gcd(c_1, \ldots, c_L)$, $\bar{c}_\ell = c_\ell/d_c$, and $\bar{B} = \lfloor B/d_c \rfloor$, the exact state recurrence $D_\ell(b)$ represents the maximum spectral energy harvested from layers $1, \ldots, \ell$ with cost at most $b$:

$$D_\ell(b) = \max_{\substack{0 \le r \le \bar{r}_\ell \\ \bar{c}_\ell r \le b}} \{D_{\ell-1}(b - \bar{c}_\ell r) + V_\ell(r)\}, \qquad b = 0, \ldots, \bar{B}, \tag{A.13}$$

initialized with $D_0(b) = 0$. Backtracking from $(L, \bar{B})$ yields the globally optimal rank profile $\boldsymbol{r}^* = (r_1^*, \ldots, r_L^*)$. The exact candidate matrix update is:

$$M_\ell^+ = K_\ell^{-1/2} U_{\ell,r_\ell^*} \Sigma_{\ell,r_\ell^*} V_{\ell,r_\ell^*}^\top C_\ell^{-1/2}. \tag{A.14}$$

### A.5 Dual-Timescale Optimization Workflow

For outer iteration $m$, the Augmented Lagrangian objective is:

$$\widehat{\Phi}_m(M) = \widehat{D}_{\mathcal{F}}(M) + \omega\widehat{S}_{\mathcal{R}}(M) + \frac{[\lambda_m + \rho(\widehat{D}_{\mathcal{R}}(M) - \varepsilon_{\mathcal{R}})]_+^2 - \lambda_m^2}{2\rho}. \tag{A.15}$$

The optimization decomposes into two alternating timescales:

**Macro Step (Subspace Refresh).** Every $\Delta t_{\mathrm{alloc}}$ iterations, we recompute ambient gradient $G$, form proposals $Z_\ell(\eta)$, solve the MCKP knapsack in Eq. (A.13), and accept candidate $M^+(\eta)$ under the Armijo condition:

$$\widehat{\Phi}_m(M^+) \le \widehat{\Phi}_m(M) - \frac{a_{\mathrm{mac}}}{2\eta}\|M^+ - M\|_\Gamma^2, \qquad a_{\mathrm{mac}} \in (0, 1). \tag{A.16}$$

Upon acceptance, we install metric-orthonormal bases $U_\ell, V_\ell$ and initialize core matrices $T_\ell = \Sigma_{\ell,r_\ell^*}$.

Table A.1: Computational complexity and storage overhead of GeSPU modules (excluding base model forward/backward passes). $S$ denotes response probes, and $r_\ell$ denotes allocated layer rank.

| **Component / Operation** | **Complexity / Storage Footprint** |
|---|---|
| Dense metric storage | $O(\sum_\ell (m_\ell^2 + n_\ell^2))$ |
| Diagonal metric storage | $O(\sum_\ell (m_\ell + n_\ell))$ |
| Dense factor inversion | One-time $O(\sum_\ell (m_\ell^3 + n_\ell^3))$ eigendecomposition |
| Diagonal factor inversion | One-time $O(\sum_\ell (m_\ell + n_\ell))$ elementwise inversion |
| Sensitivity probe calibration | $S$ vector–Jacobian products per calibration batch |
| Exact MCKP rank allocation | $O(\bar{B} \sum_\ell (\bar{r}_\ell + 1))$ per macro refresh |
| Micro core update coordinates | $\sum_\ell r_\ell^2$ trainable parameters (with frozen $U_\ell, V_\ell$) |
| **Exported LoRA Adapter** | **Exactly $\sum_\ell (m_\ell + n_\ell) r_\ell \le B$ coefficients (Zero deployment overhead)** |

**Micro Step (Isometric Core Update).** At intermediate iterations ($t \not\equiv 0 \pmod{\Delta t_{\text{alloc}}}$), bases $U_\ell, V_\ell$ remain frozen. Defining the parameter realization $\mathcal{M}_\ell(T_\ell) = K_\ell^{-1/2} U_\ell T_\ell V_\ell^\top C_\ell^{-1/2}$ and core gradient $H_\ell \coloneqq U_\ell^\top K_\ell^{-1/2} G_\ell C_\ell^{-1/2} V_\ell$, the metric displacement satisfies an exact coordinate isometry:

$$\|\mathcal{M}(T^+) - \mathcal{M}(T)\|_\Gamma^2 = \sum_{\ell=1}^{L} \|T_\ell^+ - T_\ell\|_F^2. \tag{A.17}$$

Theoretically, under exact gradient descent steps $T^+ = T - \eta_{\text{mic}} H$, certified surrogate descent is guaranteed under an Armijo condition (as proven in Appendix B.5). In practical stochastic deep learning implementations, this unconstrained Euclidean isometry critically establishes that core coordinates $T_\ell$ can be optimized directly using standard first-order adaptive optimizers (specifically AdamW; see Appendix A.7) without metric distortion or capacity violation, avoiding computationally prohibitive line searches at each training step while strictly preserving the low-rank subspace structure.

### A.6 Computational Complexity and LoRA Adapter Export

Upon outer convergence, core matrix $T_\ell$ is absorbed directly into deployable adapter factors:

$$B_\ell = K_\ell^{-1/2} U_\ell T_\ell \in \mathbb{R}^{m_\ell \times r_\ell^*}, \qquad A_\ell = C_\ell^{-1/2} V_\ell \in \mathbb{R}^{n_\ell \times r_\ell^*}, \qquad M_\ell = B_\ell A_\ell^\top. \tag{A.18}$$

Precomputed curvature factors $\{K_\ell, C_\ell\}$ and semantic prototype centroids are discarded post-training. The exported update $M_\ell = B_\ell A_\ell^\top$ strictly recovers the parameter footprint and inference latency of standard LoRA at the allocated rank profile (Table A.1).

### A.7 Implementation and Hyperparameter Details

All experiments are executed using PyTorch on NVIDIA RTX 4090 GPUs across 3 independent random seeds (reporting mean $\pm$ standard deviation). For base optimization, we employ AdamW with a base learning rate of $5 \times 10^{-4}$, $(\beta_1, \beta_2) = (0.9, 0.999)$, and weight decay $0.01$ for core matrices, utilizing PyTorch's `torch.linalg.svd` for macro spectral factorizations. For Llama-3.2-1B and 3.2-3B architectures on attention $q, v$ projections, we utilize dense K-FAC factors; for Qwen2.5-7B, we deploy the diagonal metric reduction (Eq. (A.12)) to maintain high throughput.

Curvature statistics are calibrated over $|\mathcal{R}_g| = 256$ retained sequences using $S = 32$ Monte Carlo probes, with damping coefficients set to $\delta_{h,\ell} = \delta_{y,\ell} = 10^{-3}$. The semantic anchor employs $C = 10$ prototypes with temperature $\tau = 0.1$ and loss weighting $\omega = 0.5$. The NPO suppression sensitivity is set to $\beta = 0.5$. The Augmented Lagrangian penalty parameter is initialized to $\rho = 1.0$ with optimization retention tolerance $\varepsilon_{\mathcal{R}} = 3.0 \times 10^{-3}$ on the training split $\mathcal{R}_{\text{opt}}$.

In Table 2, we explicitly distinguish test-time generalization from training constraint satisfaction: the third column (*Held-out KL*) evaluates predictive divergence on the unseen held-out test split $\mathcal{R}_{\text{test}}$ at the terminal checkpoint, whereas the relative constraint violation $v_{\max} \coloneqq \max(0, \widehat{D}_{\mathcal{R}_{\text{opt}}}(M) - \varepsilon_{\mathcal{R}})/\varepsilon_{\mathcal{R}} \times 100\%$ tracks the peak constraint violation recorded on the active optimization split $\mathcal{R}_{\text{opt}}$

across training checkpoints. Under this formulation, both full GESPU and the frozen-subspace baseline strictly satisfy the Augmented Lagrangian retention bounds throughout training ($v_{\max} = 0.0\%$), whereas uncurved or greedy rank allocations incur substantial peak infeasibility during optimization. The macro subspace refresh interval is set to $\Delta t_{\text{alloc}} = 100$ steps with line search contraction $\gamma = 0.5$. Validation checkpoints are evaluated every 20 steps, selecting the candidate that minimizes the target forget metric while preserving $\geq 95\%$ of original validation utility.

# B Proofs and Exact Spectral Allocation

## B.1 Notation and Scope

For a tuple of parameter update matrices $M = (M_1, \ldots, M_L)$, we define the ambient inner product and metric norm:

$$\langle G, H \rangle \coloneqq \sum_{\ell=1}^{L} \operatorname{tr}(G_\ell^\top H_\ell), \qquad \|H\|_\Gamma^2 \coloneqq \sum_{\ell=1}^{L} \operatorname{tr}(K_\ell H_\ell C_\ell H_\ell^\top),$$
$$\|G\|_{\Gamma,*}^2 \coloneqq \sum_{\ell=1}^{L} \|K_\ell^{-1/2} G_\ell C_\ell^{-1/2}\|_F^2. \tag{B.1}$$

All matrix square roots denote unique symmetric positive-definite roots. With $d_\ell \coloneqq \min(m_\ell, n_\ell)$, the capacity set $\mathcal{C}_B = \{M : \sum_{\ell=1}^{L} c_\ell \operatorname{rank}(M_\ell) \leq B, \operatorname{rank}(M_\ell) \leq r_\ell^{\max}\}$ is a finite union of bounded-rank determinantal varieties (non-convex constraint sets).

## B.2 Local Expansion of Predictive and Semantic Discrepancies

For categorical distribution $a$, let $F(a) \coloneqq \operatorname{diag}(a) - aa^\top$. For logit variation $u$, expanding the forward KL divergence $\mathrm{KL}(p_0 \,\|\, \operatorname{softmax}(z_0 + u))$ around $u = 0$ yields:

$$\mathrm{KL}(p_0 \| \operatorname{softmax}(z_0 + u)) = \tfrac{1}{2} u^\top F(p_0) u + o(\|u\|_2^2). \tag{B.2}$$

Applying the same expansion to soft prototype membership $\pi_0 = \operatorname{softmax}(t_0)$ and averaging over retained distribution $\mathcal{R}$, the combined discrepancy satisfies:

$$\begin{aligned} D_{\mathcal{R}}(H) + \omega S_{\mathcal{R}}(H) &= \tfrac{1}{2} \mathbb{E}_{\mathcal{R}} \left[ \dot{z}_0[H]^\top F(p_0) \dot{z}_0[H] + \omega \dot{t}_0[H]^\top F(\pi_0) \dot{t}_0[H] \right] + o(\|H\|^2) \\ &= \tfrac{1}{2} \mathbb{E}_{\mathcal{R}} \|\dot{\psi}_0[H]\|_2^2 + o(\|H\|^2). \end{aligned} \tag{B.3}$$

For layer $\ell$, directional variation is $\dot{\psi}_{0,\ell}[H] = J_{0,\ell} H_\ell h_{0,\ell}$. Kronecker-factored expectation decoupling yields $\mathbb{E}_{\mathcal{R}}[\|J_{0,\ell} H_\ell h_{0,\ell}\|_2^2] \approx \operatorname{tr}(K_\ell H_\ell C_\ell H_\ell^\top)$, establishing the retain metric.

## B.3 Proof of Theorem 1 (Exact Spectral Projection)

Let $Y_\ell \coloneqq \Gamma_\ell(M_\ell) = K_\ell^{1/2} M_\ell C_\ell^{1/2}$, $Y_{\ell,t} \coloneqq \Gamma_\ell(M_{\ell,t})$, and transformed gradient $\widetilde{G}_{\ell,t} = K_\ell^{-1/2} G_{\ell,t} C_\ell^{-1/2}$. Setting proposal $Z_{\ell,t} = Y_{\ell,t} - \eta \widetilde{G}_{\ell,t}$, completing the square in surrogate Eq. (10) yields:

$$Q_t(M) = \frac{1}{2\eta} \sum_{\ell=1}^{L} \|Y_\ell - Z_{\ell,t}\|_F^2 - \frac{\eta}{2} \sum_{\ell=1}^{L} \|\widetilde{G}_{\ell,t}\|_F^2. \tag{B.4}$$

Because factors $K_\ell, C_\ell$ are non-singular, coordinate transformation $\Gamma_\ell$ preserves rank identically: $\operatorname{rank}(Y_\ell) = \operatorname{rank}(M_\ell)$. For any fixed layerwise rank assignment $(r_1, \ldots, r_L)$, the Eckart–Young–Mirsky theorem (Eckart & Young, 1936) states that truncated SVD $[Z_{\ell,t}]_{r_\ell}$ minimizes each Frobenius distance over rank-$r_\ell$ matrices (uniquely so when $\sigma_{\ell,r_\ell,t} > \sigma_{\ell,r_\ell+1,t}$, with any tied spectral subspace achieving the identical global surrogate value):

$$\min_{\operatorname{rank}(Y_\ell) \leq r_\ell} \|Y_\ell - Z_{\ell,t}\|_F^2 = \sum_{i=r_\ell+1}^{d_\ell} \sigma_{\ell,i,t}^2 = \|Z_{\ell,t}\|_F^2 - V_{\ell,t}(r_\ell), \tag{B.5}$$

where $V_{\ell,t}(r) = \sum_{i=1}^{r} \sigma_{\ell,i,t}^2$. Substituting into Eq. (B.4) reduces surrogate minimization to maximizing cumulative spectral energy $\sum_\ell V_{\ell,t}(r_\ell)$ subject to $\sum_\ell c_\ell r_\ell \le B$, which is an exact Multiple-Choice Knapsack Problem (MCKP). Inverting the coordinate map yields global minimizer $M_{\ell,t}^+ = K_\ell^{-1/2}[Z_{\ell,t}]_{r_\ell^*} C_\ell^{-1/2}$. □

### B.4 Discrete Concavity and Knapsack Complexity

Dynamic programming recurrence Eq. (A.13) over state space $\bar{B} = \lfloor B/\gcd(c_1, \dots, c_L) \rfloor$ solves the MCKP in $O(\bar{B}\sum_\ell r_\ell^{\max})$ operations. Furthermore, because singular values are ordered $\sigma_{\ell,1,t} \ge \sigma_{\ell,2,t} \ge \dots$, marginal returns $V_\ell(r) - V_\ell(r-1) = \sigma_{\ell,r,t}^2$ are non-increasing, meaning $V_\ell(r)$ is discrete-concave. By residue splitting $b = s{+}cj$, the decision matrix $A_{jk} = D_{\ell-1}(s{+}ck){+}V_\ell(j{-}k)$ satisfies the anti-Monge property:

$$A_{j_1k_1} + A_{j_2k_2} \ge A_{j_1k_2} + A_{j_2k_1}, \qquad \forall j_1 < j_2,\ k_1 < k_2. \tag{B.6}$$

Consequently, optimal column transitions are monotonic, enabling a divide-and-conquer acceleration that solves each layer in $O(N_s \log(r_\ell^{\max} + 2))$, reducing total complexity to $O\big((\bar{B} + 1)\sum_{\ell=1}^{L} \log(r_\ell^{\max} + 2)\big)$.

### B.5 Proof of Strict Descent under Backtracking

Let $f = \Phi_m$. Assuming $L_\Gamma$-Lipschitz smoothness of $\nabla f$ on a local neighborhood under metric norm $\|\cdot\|_\Gamma$:

$$f(M_t + H) \le f(M_t) + \langle G_t, H\rangle + \frac{L_\Gamma}{2}\|H\|_\Gamma^2. \tag{B.7}$$

For exact surrogate minimizer displacement $\Delta_\eta \coloneqq M_t^+(\eta) - M_t$, optimality over feasible iterate $M_t$ implies $\langle G_t, \Delta_\eta\rangle \le -\frac{1}{2\eta}\|\Delta_\eta\|_\Gamma^2$. Substituting into smoothness bound Eq. (B.7) yields:

$$f(M_t^+(\eta)) \le f(M_t) - \left(\frac{1}{2\eta} - \frac{L_\Gamma}{2}\right)\|\Delta_\eta\|_\Gamma^2. \tag{B.8}$$

For any acceptance constant $a \in (0,1)$, whenever step size satisfies $\eta \le (1-a)/L_\Gamma$, the Armijo decrease condition $f(M_t^+(\eta)) \le f(M_t) - \frac{a}{2\eta}\|\Delta_\eta\|_\Gamma^2$ is strictly satisfied. Therefore, geometric backtracking $\eta \leftarrow \gamma\eta$ $(0 < \gamma < 1)$ terminates in a finite number of trials, guaranteeing strict descent for non-zero displacements. □

### B.6 Proof of Corollary 1 (Spectral Steering at Initialization)

At initial state $M_0 = 0$ with $\lambda_0 = 0$, base predictions match targets identically ($p_{M=0} = p_0, \pi_{M=0} = \pi_0$), causing retention gradients to vanish ($\nabla D_{\mathcal{R}}(0) = 0, \nabla S_{\mathcal{R}}(0) = 0$). The ambient gradient reduces strictly to $G_{\ell,0} = \nabla_{M_\ell} D_{\mathcal{F}}(0)$, and the proposal simplifies to $Z_{\ell,0} = -\eta K_\ell^{-1/2}[\nabla_{M_\ell} D_{\mathcal{F}}(0)] C_\ell^{-1/2}$. Applying Theorem 1 directly yields the closed-form energy harvested at step zero:

$$-\min_{M \in \mathcal{C}_B} Q_0(M) = \frac{\eta}{2} \max_{\sum c_\ell r_\ell \le B} \sum_{\ell=1}^{L}\sum_{i=1}^{r_\ell} \sigma_i^2\Big(K_\ell^{-1/2}\big[\nabla_{M_\ell} D_{\mathcal{F}}(0)\big] C_\ell^{-1/2}\Big). \tag{B.9}$$

Spectral preconditioning by $K_\ell^{-1/2}$ and $C_\ell^{-1/2}$ scales coordinates inversely by their retain curvature $\kappa_i\chi_j$, attenuating sensitive directions prior to truncation. □

### B.7 Core-Coordinate Isometry and LoRA Export

Given metric-orthonormal bases $U_\ell^\top U_\ell = I, V_\ell^\top V_\ell = I$, parameter updates parameterized by core matrices $T_\ell, T_\ell' \in \mathbb{R}^{r_\ell^* \times r_\ell^*}$ satisfy $M_\ell(T_\ell) = K_\ell^{-1/2} U_\ell T_\ell V_\ell^\top C_\ell^{-1/2}$. By the trace cycle property:

$$\begin{aligned}\|M_\ell(T_\ell) - M_\ell(T_\ell')\|_\Gamma^2 &= \operatorname{tr}\Big((T_\ell - T_\ell')^\top U_\ell^\top K_\ell^{-1/2} K_\ell K_\ell^{-1/2} U_\ell \\ &\qquad \times (T_\ell - T_\ell') V_\ell^\top C_\ell^{-1/2} C_\ell C_\ell^{-1/2} V_\ell\Big) \\ &= \|T_\ell - T_\ell'\|_F^2.\end{aligned} \tag{B.10}$$

This establishes an exact coordinate isometry between the Riemannian metric displacement and Euclidean core distances. Finally, defining factors $B_\ell \coloneqq K_\ell^{-1/2} U_\ell T_\ell$ and $A_\ell \coloneqq C_\ell^{-1/2} V_\ell$ recovers deployable LoRA updates $M_\ell = B_\ell A_\ell^\top$ with zero added inference overhead. □