FROM WHAT IS USED TO WHAT IS MISSING: RETHINKING DYNAMIC RANK IN LORA

Anonymous authors
Paper under double-blind review

ABSTRACT

Parameter-efficient fine-tuning with LoRA is governed by rank, which controls parameter budget and locally accessible task updates. Existing dynamic-rank methods mainly focus on capacity allocation, leaving unresolved whether the current adaptation space is sufficient for the task. We characterize the first-order information boundary of standard factor gradients, showing that their unobservable subspace coincides with the normal space of the low-rank manifold. Thus, factor gradients alone cannot distinguish under-optimization within the current space from task-relevant directions beyond its first-order reach. This motivates a geometric view of rank evolution, where tangent information optimizes existing degrees of freedom and the normal spectrum quantifies the value of new ones. Under a local quadratic model, we show that leading normal singular directions yield the locally optimal rank expansion, with spectral energy determining its gain. Building on this characterization, we propose GEAR, which turns the tangent–normal decomposition into rank evolution through tangent-space optimization, matrix-free normal-spectrum estimation for rank expansion, and task-induced geometric initialization. Across language and vision benchmarks, GEAR achieves the best average performance in all evaluated settings with substantially fewer trainable parameters. On LLaMA3-8B, it raises the average commonsense score from 86.61 to 88.45 while reducing trainable parameters from 42.0M to 14.4M.

1 INTRODUCTION

Parameter-efficient fine-tuning (PEFT) has become a prevailing approach to adapting large pretrained models by updating only a small fraction of task-specific parameters (Ding et al., 2023b). Among PEFT methods, LoRA represents the weight increment as a low-rank matrix, $\Delta W = BA$, where $A \in \mathbb{R}^{r\times m}$ and $B \in \mathbb{R}^{n\times r}$, reducing the number of trainable parameters from nm to $r(n+m)$ (Hu et al., 2022). While r therefore directly controls the adaptation budget, it also determines the local first-order update space accessible under the low-rank constraint (Vandereycken, 2013). Rank thus governs both adaptation capacity and which task-induced directions can be expressed. Since these requirements vary across layers and throughout training, a uniform fixed rank is unlikely to match the evolving structure of task demand, motivating dynamic-rank adaptation.

Most dynamic-rank methods still treat rank primarily as a capacity-allocation variable. AdaLoRA, for example, reallocates a prescribed rank budget across weight matrices according to the importance of individual singular components (Zhang et al., 2023), while DyLoRA trains nested low-rank structures to support inference at multiple truncation ranks (Valipour et al., 2023). More recent approaches permit more flexible rank evolution: FlexLoRA uses the spectral entropy of the current low-rank update to determine when capacity should be pruned or expanded (Liu et al., 2026), whereas GeoLoRA combines gradient-driven basis augmentation with spectral truncation within a dynamical low-rank integration framework (Schotthöfer et al., 2025). Although these methods substantially improve the flexibility of rank adaptation, their rank decisions are primarily derived from the capacity or structure already represented by the current adapter. They therefore characterize what the current low-rank space uses, but provide no direct criterion for determining whether task-relevant descent directions remain missing outside that space. This leaves a more fundamental question unresolved:

Is the current adaptation space locally sufficient to express task-relevant descent directions, or must its rank expand to make such directions first-order accessible?

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.

Answering this question requires distinguishing insufficient optimization within the current low-rank space from insufficiency of the space itself. Limited progress may arise for two fundamentally different reasons: the current rank may be sufficient but under-optimized, or task-relevant descent directions may lie outside its first-order reachable space. Only the latter justifies increasing rank. Existing rank-allocation criteria do not explicitly separate these two cases, making it difficult to determine whether additional capacity is genuinely required. This distinction motivates our central view that dynamic rank should be driven not only by what the current adaptation space already represents, but also by what task-relevant descent information remains inaccessible to it.

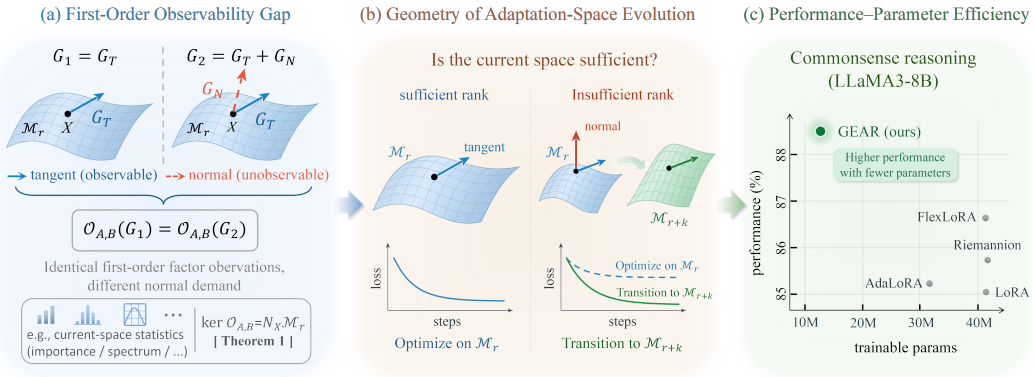

Figure 1: Motivation and overview of GEAR. (a) standard factor-gradient observations are insensitive to normal components, so **identical first-order observations can correspond to different rank requirements**. (b) GEAR distinguishes optimization within the current manifold from expansion toward task-relevant directions beyond the current first-order reachable space. (c) this geometry-guided rank evolution yields a stronger performance–parameter trade-off on commonsense reasoning.

To address this limitation, we propose GEAR (Geometry-guided Evolution of Adaptive Rank), which formulates dynamic-rank adaptation as the evolution of a task-dependent low-rank space. Rather than inferring rank adequacy from existing factor statistics, we characterize the first-order information available under the fixed-rank parameterization. For $X = BA$, we show that the standard factor-gradient map has an observability gap whose null space coincides with the normal space $N_X\mathcal{M}_r$ of the fixed-rank manifold $\mathcal{M}_r = \{X \in \mathbb{R}^{n\times m} : \mathrm{rank}(X) = r\}$. This establishes a structural information limitation in which standard factor gradients reveal the tangent component relevant to optimizing the current rank, while leaving the normal component unobservable. Consequently, they cannot distinguish under-optimization within the current adaptation space from task-relevant directions requiring rank expansion to become first-order accessible. For a prescribed rank increase, we further show that leading normal modes determine the locally optimal expansion, while their spectral energy quantifies its gain. Building on this characterization, we propose GEAR, which turns the tangent–normal decomposition into a practical rank-evolution mechanism through tangent-space optimization, matrix-free normal-spectrum estimation for rank expansion, and task-induced geometric initialization. By expanding rank only when the estimated normal spectrum reveals valuable task-relevant directions, GEAR improves adaptation effectiveness and parameter efficiency. Across multiple benchmarks, it consistently outperforms competitive PEFT baselines while using fewer trainable parameters. Our contributions are as follows:

- **First-order observability barrier for rank adaptation.** To our knowledge, we are the first to characterize the information available from standard fixed-rank factor gradients and prove that their unobservable subspace is the normal space of the low-rank manifold, thereby separating current-space optimization from rank-expanding directions.
- **Geometry-guided rank evolution.** We introduce GEAR, which unifies optimization over the current fixed-rank manifold and rank expansion using tangent information to optimize the adaptation space and the normal spectrum to identify and value new task-relevant directions.
- **Excellent performance–parameter trade-off.** Across language and vision benchmarks, GEAR achieves the best average performance in every evaluated setting while using substantially fewer trainable parameters than strong PEFT baselines.

2 Related work

LoRA and Adaptive Rank Adaptation. LoRA parameterizes task-specific updates through low-rank factors, reducing trainable parameters and storage while preserving frozen pretrained weights (Hu et al., 2022). Subsequent work has improved low-rank adaptation through refined parameterizations and initialization strategies. DoRA decouples weight magnitude and direction to enhance adaptation flexibility (Liu et al., 2024), while PiSSA initializes adapters from principal singular subspaces of the pretrained weights (Meng et al., 2024) and LoRA-GA uses task gradients to identify more effective initial low-rank directions (Wang et al., 2024). Despite these advances, most methods still operate with a fixed rank, which may poorly match the heterogeneous adaptation demands across layers and training stages. This limitation has motivated adaptive-rank LoRA methods. AdaLoRA dynamically reallocates a prescribed rank budget across modules according to singular-component importance (Zhang et al., 2023); DyLoRA trains nested low-rank structures to support multiple effective ranks (Valipour et al., 2023); and SoRA learns effective rank through sparse gating (Ding et al., 2023a). More recent approaches allow rank to both contract and expand: FlexLoRA adjusts rank according to the spectral structure of the current adapter (Liu et al., 2026), whereas GeoLoRA combines gradient-driven basis augmentation with spectral truncation under a dynamical low-rank formulation (Schotthöfer et al., 2025). GeLoRA further estimates the intrinsic dimensionality of hidden representations to derive adaptive layer-wise ranks (Ed-dib et al., 2025). Collectively, these methods have advanced LoRA from fixed-rank adaptation toward increasingly flexible, task-dependent capacity allocation. However, existing adaptive-rank methods primarily decide how to allocate or reorganize represented low-rank capacity, whereas GEAR asks a different question: whether task-relevant descent information remains outside the current first-order reachable space and therefore warrants new rank dimensions.

Geometric Optimization for LoRA Fixed-rank matrices admit a natural Riemannian manifold structure, providing a principled way to optimize low-rank states independently of a particular factorization (Vandereycken, 2013). Early geometric approaches to LoRA exploit this structure through Riemannian metrics and preconditioning; for example, Riemannian Preconditioned LoRA derives low-dimensional preconditioners from the geometry of the low-rank matrix space to improve optimization stability and convergence (Zhang & Pilanci, 2024). GeoLoRA further introduces dynamical low-rank integration into PEFT, combining gradient-driven basis augmentation with spectral truncation for structure-preserving and rank-adaptive updates (Schotthöfer et al., 2025). Related work has also explored explicit manifold constraints on LoRA factors. Park et al. optimize the B factor on the Stiefel manifold to maintain orthogonality and effective rank (Park et al., 2025), while Manifold-LoRA develops a retraction-free Stiefel-manifold formulation for LoRA fine-tuning (Zhang et al., 2026). More recently, Riemannion directly optimizes the adaptation matrix on the fixed-rank manifold, removing factorization ambiguity and enabling parametrization-independent geometric updates (Bogachev et al., 2026).

These methods establish geometry as an effective tool for improving low-rank optimization, primarily by refining the metric, constraining the factor geometry, or evolving the current low-rank state. GEAR instead focuses on the information required to determine when the current adaptation space itself should expand. From a first-order observability perspective, it relates the normal component unobservable from standard factor gradients to task-relevant directions beyond the current first-order reachable space, thereby connecting manifold optimization with rank expansion.

3 Method

GEAR builds on a first-order observability gap in fixed-rank adaptation: standard factor gradients characterize optimization within the current space, but not the normal task information that governs its expansion. We first formalize this boundary and then develop a geometry-guided adaptation framework that optimizes the low-rank space while identifying when and how it should expand.

3.1 First-Order Information Boundary of Standard Factor Gradients

Consider a pretrained weight matrix $W \in \mathbb{R}^{n\times m}$ adapted as

$$W' = W + sX, \qquad X = BA,$$

Algorithm 1: GEAR: Geometry-Guided Evolution of Adaptive Rank

Input: Pretrained weights W; initial rank r_0; rank upper bound $r_{\max}$; audit set $\mathcal{T}_{\text{audit}}$; budget $\mathcal{B}_{\text{target}}$

$(U, C, V) \leftarrow \text{TGI}(W, r_0)$;
$X \leftarrow UC, \quad W \leftarrow W - sX, \quad r \leftarrow r_0$;
for $t = 1, \ldots, T$ **do**
  $G_T \leftarrow U(U^\top G) + (I - UU^\top)GVV^\top$; // Eq. (1)
  $(X, U, C, V) \leftarrow \text{Retr}_{\mathcal{M}_r}(X - \eta_t G_T)$; // Eq. (2)
  if $t \in \mathcal{T}_{\text{audit}}$ ***and*** $r < r_{\max}$ **then**
    $(\widehat{U}, \widehat{\Sigma}, \widehat{V}) \leftarrow \text{NormalSketch}(X)$; // Eqs. (3)–(4)
    $\mathcal{S} \leftarrow \text{Select}(\widehat{U}, \widehat{\Sigma}, \widehat{V}; \mathcal{B}_{\text{target}})$; // Eq. (5)
    $X_{\text{cand}} \leftarrow \text{Propose}(X, \mathcal{S})$; // Eq. (6)
    if $\text{Validate}(X_{\text{cand}}, X)$ **then**
      $(U, C, r) \leftarrow \text{Expand}(U, C, \mathcal{S}), \quad X \leftarrow UC$; // Eq. (7)
    end
  end
end
Output: X

where $s > 0$ is a fixed scaling factor. Let $f(X) = \mathcal{L}(W + sX)$ and $G = \nabla_X f(X)$. For a rank-r adaptation matrix $X = BA$, we analyze X on the fixed-rank manifold $\mathcal{M}_r$, whose tangent space characterizes the first-order directions accessible without increasing rank (Vandereycken, 2013).

Let $X = U\Sigma V^\top$ be the compact SVD of X, and define $P_U = UU^\top$ and $P_V = VV^\top$. The full gradient admits the orthogonal decomposition

$$G = G_T + G_N,$$

where $G_T = P_U G + (I - P_U)GP_V \in T_X\mathcal{M}_r$ and $G_N = (I - P_U)G(I - P_V) \in N_X\mathcal{M}_r$. Under the factorized parameterization $X = BA$, the corresponding gradients are

$$\nabla_B f = GA^\top, \qquad \nabla_A f = B^\top G.$$

We therefore define the factor-gradient observation operator

$$\mathcal{O}_{A,B}(G) = \big(GA^\top, B^\top G\big),$$

which captures the first-order information about G available from standard factor backpropagation.

Theorem 1 (Factor-Gradient Observability Barrier)**.** *Let $A \in \mathbb{R}^{r\times m}$ and $B \in \mathbb{R}^{n\times r}$ be full rank, with $X = BA \in \mathcal{M}_r$. Then*

$$\ker \mathcal{O}_{A,B} = N_X\mathcal{M}_r.$$

Equivalently, for any G_1 and G_2,

$$\mathcal{O}_{A,B}(G_1) = \mathcal{O}_{A,B}(G_2)$$

if and only if $(G_1)_T = (G_2)_T$.

The theorem establishes a precise first-order observability boundary, showing that standard factor gradients preserve the tangent component of the task gradient but cannot distinguish gradients that differ only in their normal components. A complete proof is provided in Appendix C.1. Consequently, the same factor-gradient observation may correspond either to a gradient fully represented by the current adaptation space or to one containing additional task-relevant directions beyond its first-order reach. This ambiguity arises from the information available under the current fixed-rank parameterization, rather than from factorization itself.

This result distinguishes optimization within the current adaptation space from the need for rank expansion. The tangent component represents directions that are first-order accessible at rank r, whereas the normal component captures task-relevant directions beyond the current space. Accordingly, tangent information is used to optimize the existing degrees of freedom, while normal information provides the relevant signal for assessing the value of rank expansion. The subsequent sections develop these two mechanisms, respectively.

3.2 Fixed-Rank Manifold Optimization

For a fixed rank r, GEAR performs optimization on the manifold $\mathcal{M}_r$ using the tangent component of the gradient. We represent the adaptation matrix as

$$X = UC, \qquad U^\top U = I_r,$$

where U is an orthonormal basis for the column space of X, and C contains the corresponding low-rank coefficients. Let V be an orthonormal basis for the column space of $C^\top$. The tangent gradient is then

$$G_T = U(U^\top G) + (I - UU^\top)GVV^\top. \tag{1}$$

Equation (1) shows that G_T can be computed from the two gradient contractions $U^\top G$ and GV, avoiding explicit construction of the full gradient G.

To obtain these quantities directly through backpropagation, GEAR uses the forward-equivalent representation

$$B_{\text{ext}} = [U,\ 0], \qquad A_{\text{ext}} = \begin{bmatrix} C \\ V^\top \end{bmatrix},$$

which satisfies $B_{\text{ext}} A_{\text{ext}} = X$. Although this representation leaves the forward mapping unchanged, its backward derivatives expose the required gradient contractions:

$$\nabla_C f = U^\top G, \qquad \nabla_{B_{\text{ext}}^{(2)}} f = GV.$$

Thus, the tangent gradient can be recovered without explicitly materializing G. GEAR then performs a tangent update followed by retraction onto the fixed-rank manifold,

$$X^- = X - \eta G_T, \qquad X_{t+1} = \text{Retr}_{\mathcal{M}_r}(X^-). \tag{2}$$

Since X^- has rank at most $2r$, the retraction is performed on its compact factors rather than on the full matrix. Specifically, thin QR factorizations reduce the update to a small core matrix, whose truncated SVD yields the retracted rank-r state together with the updated left and right subspace bases. Full details are provided in Appendix B.1.

3.3 Geometry-Guided Rank Evolution

The preceding section optimizes the current adaptation state on $\mathcal{M}_r$. Rank expansion, by contrast, is governed by the normal component G_N, which captures task-relevant directions beyond the current first-order reachable space. We next quantify the local value of these directions for rank expansion. Let

$$G_N = \sum_{i \geq 1} \sigma_i u_i v_i^\top, \qquad \sigma_1 \geq \sigma_2 \geq \cdots,$$

and denote its best rank-k approximation by $(G_N)_k$.

Theorem 2 (Local Optimality of Rank Expansion)**.** *Let $X \in \mathcal{M}_r$, and let $\mathcal{C}_{r,k}(X)$ denote the tangent cone at X to the bounded-rank set $\mathcal{V}_{r+k} = \{Y \in \mathbb{R}^{n\times m} : \text{rank}(Y) \leq r+k\}$. For the local quadratic model $q_\eta(H) = \langle G, H\rangle_F + \frac{1}{2\eta}\|H\|_F^2$, an optimal direction over $\mathcal{C}_{r,k}(X)$ is*

$$H_k^\star = -\eta\,[G_T + (G_N)_k]\,,$$

where $(G_N)_k$ is a best rank-k approximation of G_N. Moreover, the optimal model decrease enabled by the additional k rank dimensions, relative to optimization over $T_X\mathcal{M}_r$, is

$$\Delta_k := \min_{H \in T_X\mathcal{M}_r} q_\eta(H) - \min_{H \in \mathcal{C}_{r,k}(X)} q_\eta(H) = \frac{\eta}{2}\sum_{i=1}^{k} \sigma_i^2(G_N).$$

A complete proof is provided in Appendix C.2. The result shows that the leading singular directions of the normal gradient constitute the locally optimal directions for rank expansion, while their spectral energy exactly quantifies the additional decrease attainable under the local quadratic model.

However, constructing $G_N = (I - UU^\top)G(I - VV^\top)$ requires the full gradient G. GEAR instead estimates its dominant spectrum through randomized probing (Halko et al., 2011). Given

$\Omega \in \mathbb{R}^{m\times h}$, with $h \ll \min\{n, m\}$, we project it onto the orthogonal complement of the right subspace, $\Omega_\perp = (I - VV^\top)\Omega$. We introduce a zero-initialized auxiliary variable P through $\Delta X_{\text{probe}} = P\Omega_\perp^\top$. Since $P = 0$, this leaves the forward mapping unchanged, while backpropagation yields $\nabla_P f = G\Omega_\perp$. Projecting onto the orthogonal complement of the left subspace gives

$$Y = (I - UU^\top)\nabla_P f = G_N\Omega. \tag{3}$$

Let $Q = \text{orth}(Y)$. If $G_N = U_N\Sigma_N V_N^\top$, then $Y = U_N\Sigma_N(V_N^\top\Omega)$, so singular directions associated with larger singular values contribute more strongly to Y. Hence, $\text{span}(Q)$ captures the dominant left singular subspace of G_N. Applying the same zero-output probe yields $Q^\top G$, from which we compute

$$Z_N = Q^\top G(I - VV^\top) = Q^\top G_N = \widetilde{U}\widehat{\Sigma}\widehat{V}^\top. \tag{4}$$

The leading normal components are therefore approximated by $\widehat{u}_i = Q\widetilde{u}_i$, $\widehat{v}_i = \widehat{V}_{:,i}$, and $\widehat{\sigma}_i = \widehat{\Sigma}_{ii}$, without explicitly materializing G_N. The resulting approximation has a controlled effect on the local value of rank expansion.

Theorem 3 (Stability under Inexact Normal Estimation). *Let $\widehat{G}_N \in N_X\mathcal{M}_r$ satisfy $\|\widehat{G}_N - G_N\|_F \le \varepsilon$, and let $\tau_k = \|G_N - (G_N)_k\|_F$. Then the rank-k expansion based on $(\widehat{G}_N)_k$ satisfies*

$$0 \le q_\eta\Big(-\eta(\widehat{G}_N)_k\Big) - q_\eta(-\eta(G_N)_k) \le 2\eta\varepsilon(\tau_k + \varepsilon).$$

A complete proof is provided in Appendix C.3. GEAR then uses the estimated normal spectrum to allocate rank under a global parameter budget. Motivated by Theorem 2, we score each candidate normal direction by its estimated local gain per added parameter and impose a global rank budget:

$$a_{\ell,i} = \frac{\widehat{\sigma}_{\ell,i}^2}{n_\ell + m_\ell}, \qquad \mathcal{B}_{\text{target}} = \bar{r}_{\text{budget}} \sum_{\ell=1}^{L} (n_\ell + m_\ell). \tag{5}$$

Here, r_ℓ denotes the current rank of module ℓ, $n_\ell + m_\ell$ is the cost of one additional rank dimension in module ℓ, and $\bar{r}_{\text{budget}}$ denotes the weighted-average rank budget. Candidates are selected according to $a_{\ell,i}$ subject to the global budget and the per-module rank limit. For k_ℓ selected directions, let $\widehat{U}_{\ell,k_\ell} = [\widehat{u}_{\ell,1}, \ldots, \widehat{u}_{\ell,k_\ell}]$, $\widehat{V}_{\ell,k_\ell} = [\widehat{v}_{\ell,1}, \ldots, \widehat{v}_{\ell,k_\ell}]$, and $\widehat{\Sigma}_{\ell,k_\ell} = \text{diag}(\widehat{\sigma}_{\ell,1}, \ldots, \widehat{\sigma}_{\ell,k_\ell})$. They define the candidate expansion

$$X_{\ell,\text{cand}} = X_\ell - \eta_b\widehat{U}_{\ell,k_\ell}\widehat{\Sigma}_{\ell,k_\ell}\widehat{V}_{\ell,k_\ell}^\top. \tag{6}$$

We validate $X_{\ell,\text{cand}}$ on a small set of separate training microbatches shared across candidate variants and accept the expansion only if its mean loss improves by more than a relative margin ρ; full details are provided in Appendix B.2.

For an accepted expansion, $\widehat{U}_{\ell,k_\ell} \perp U_\ell$ and $\widehat{V}_{\ell,k_\ell} \perp V_\ell$, so the low-rank state is updated directly as

$$U_\ell^+ = [U_\ell, \widehat{U}_{\ell,k_\ell}], \qquad C_\ell^+ = \begin{bmatrix} C_\ell \\ -\eta_b\widehat{\Sigma}_{\ell,k_\ell}\widehat{V}_{\ell,k_\ell}^\top \end{bmatrix}, \qquad X_\ell^+ = U_\ell^+ C_\ell^+. \tag{7}$$

The accepted expansion moves the adaptation state from $\mathcal{M}_{r_\ell}$ to $\mathcal{M}_{r_\ell + k_\ell}$. Optimizer states associated with existing components are inherited, while those of newly introduced directions are initialized to zero. Subsequent updates then follow the fixed-rank manifold optimization in Section 3.2.

3.4 Task-Induced Geometric Initialization (TGI)

Rank inadequacy is defined relative to the current adaptation space. To avoid conflating random subspace mismatch with genuine missing directions, GEAR initializes this space from task-gradient structure estimated on a small calibration set. Let $G_0 = \nabla_X f(0)$. In a left-probing stage, set $X = P\Omega^\top$ with $P = 0$ and fixed $\Omega \in \mathbb{R}^{m\times h}$. This preserves the pretrained mapping, while backpropagation yields $Z_0 = \nabla_P f(P\Omega^\top)\big|_{P=0} = G_0\Omega \in \mathbb{R}^{n\times h}$ without forming G_0.

To express the normalization for both stages uniformly, write $Z_0 \in \mathbb{R}^{d\times h}$, with $d = n$ for the left probe and $d = m$ for the right probe. Set $M = Z_0^{\odot 2}$ and $\mu = (dh)^{-1}\sum_{i,a} M_{ia}$. To prevent large gradient scales from dominating subspace estimation, we normalize the sketch along both dimensions:

$$d_i^{\text{row}} = \frac{1}{h}\sum_a M_{ia} + \lambda\mu, \quad d_a^{\text{col}} = \frac{1}{d}\sum_i M_{ia} + \lambda\mu, \quad \widetilde{Z}_0 = D_{\text{row}}^{-1/2} Z_0 D_{\text{col}}^{-1/2},$$

where $D_{\text{row}} = \text{diag}(d_i^{\text{row}})$ and $D_{\text{col}} = \text{diag}(d_a^{\text{col}})$. A thin QR factorization of the left-stage $\widetilde{Z}_0$ yields $Q_L \in \mathbb{R}^{n\times h}$. Holding Q_L fixed, a zero-initialized right probe $X = Q_L R$ gives the transposed sketch $Z_0 = (\nabla_R f(Q_L R)|_{R=0})^\top = G_0^\top Q_L \in \mathbb{R}^{m\times h}$. Non-final right stages use QR to update the next left-probing basis. TGI performs $p+1$ left–right rounds, totaling $2(p+1)$ stages. Additional rounds can refine the dominant-subspace estimate by probing with successively updated bases; we use $p = 1$ in our experiments.

At the final right stage, the normalized sketch $\widetilde{Z}_0 \in \mathbb{R}^{m\times h}$ gives $\bar{Z}_0 = \widetilde{Z}_0^\top = \widehat{U}\Sigma V^\top$. Retaining the leading $r_0 \le h$ components, we initialize

$$U_0 = Q_L \widehat{U}_{:,1:r_0}, \qquad C_0 = \text{diag}(\gamma_1, \ldots, \gamma_{r_0}) V^\top_{:,1:r_0}, \qquad X_0 = U_0 C_0. \tag{8}$$

Here γ_i is the clipped ratio of a retained singular value to the mean of the retained values. Since X_0 is generally nonzero, we compensate its initial contribution in the frozen weight:

$$\widetilde{W} = W - sX_0, \qquad \widetilde{W} + sX_0 = W.$$

Thus, training starts from a task-aligned adaptation space while preserving the pretrained forward mapping at initialization.

4 Experimental results

We evaluate GEAR across commonsense reasoning, natural language understanding, and visual recognition benchmarks, considering both predictive performance and parameter efficiency. We compare against representative fixed-rank and adaptive-rank PEFT methods, including LoRA (Hu et al., 2022), AdaLoRA (Zhang et al., 2023), Riemannion (Bogachev et al., 2026), GeoLoRA (Schotthöfer et al., 2025), and FlexLoRA (Liu et al., 2026). All methods use consistent experimental settings for fair comparison. Ablation studies are provided in Appendix E.

4.1 Rank Audit Schedule

To ensure stable rank adaptation, GEAR performs rank-growth audits periodically during early-to-middle training and freezes the rank thereafter. Let t_{start} denote the first audit step, Δt the audit interval, and t_{final} the step after which rank expansion stops. The audit steps are defined as

$$\mathcal{T}_{\text{audit}} = \{t_{\text{start}} + k\Delta t \mid k \in \mathbb{Z}_{\ge 0},\ t_{\text{start}} + k\Delta t \le t_{\text{final}}\}.$$

At each audit, GEAR estimates task-relevant directions beyond the current adaptation space and evaluates them as candidates for rank expansion. After t_{final}, the rank remains fixed and training proceeds solely with manifold optimization at the resulting rank.

4.2 Commonsense Reasoning

We evaluate GEAR on eight commonsense reasoning benchmarks, including BoolQ, PIQA, SIQA, and HellaSwag (Clark et al., 2019; Bisk et al., 2020; Sap et al., 2019; Zellers et al., 2019), together with WinoGrande, ARC-e/ARC-c, and OBQA (Sakaguchi et al., 2020; Clark et al., 2018; Mihaylov et al., 2018), using LLaMA3-8B (Grattafiori et al., 2024) and Qwen2.5-7B (Qwen et al., 2025) as backbone models. The reference rank is set to 16, while GEAR initializes at $r_0 = 4$, with a weighted-average rank budget of 16 and a per-module upper bound $r_{\max} = 32$; the final ranks are determined by the accepted expansions rather than the upper bound. Performance is reported together with the number of trainable parameters to evaluate both adaptation quality and parameter efficiency.

As shown in Table 1, GEAR achieves the highest average performance on both backbones with substantially fewer trainable parameters. On LLaMA3-8B, it reaches 88.45, outperforming FlexLoRA by 1.84 points with 14.4M versus 42.0M parameters; on Qwen2.5-7B, it attains 89.40 with 14.0M parameters, versus 87.65 with 40.3M for FlexLoRA. Unlike adaptive-rank methods that rely mainly on statistics of the current low-rank representation, GEAR expands the adaptation space according to task-relevant normal directions while optimizing accessible directions on the current manifold, yielding a better performance–parameter trade-off across both backbones.

Table 1: Commonsense reasoning performance on LLaMA3-8B and Qwen2.5-7B. We report the mean over three runs with different random seeds. Best results are shown in **bold**.

Method (Params)	BoolQ	PIQA	SIQA	HellaS.	WinoG.	ARC-e	ARC-c	OBQA	Avg.
LLaMA3-8B									
LoRA (41.9M)	72.37	87.83	81.34	94.81	85.08	91.46	81.32	86.00	85.03
AdaLoRA (31.6M)	72.23	87.21	81.27	94.60	85.16	93.18	82.68	85.80	85.27
Riemannion (41.9M)	72.45	87.32	81.17	94.95	87.64	93.14	82.51	87.40	85.82
GeoLoRA (41.3M)	72.84	87.54	79.99	94.62	85.56	93.52	83.96	87.00	85.63
FlexLoRA (42.0M)	72.91	88.30	82.96	95.68	87.14	93.56	83.70	88.60	86.61
GEAR (**14.4M**)	**75.19**	**90.81**	**83.78**	**96.82**	**89.82**	**93.77**	**85.58**	**91.80**	**88.45**
Qwen2.5-7B									
LoRA (40.4M)	71.54	88.52	78.73	94.05	82.56	96.13	90.00	90.12	86.46
AdaLoRA (35.2M)	71.59	89.12	78.92	94.06	82.79	96.00	90.27	89.80	86.57
Riemannion (40.4M)	71.68	89.17	79.32	94.72	84.53	95.96	**90.53**	91.00	87.11
GeoLoRA (40.1M)	71.90	89.12	79.02	94.88	85.00	96.21	89.93	91.00	87.13
FlexLoRA (40.3M)	72.29	89.17	79.68	95.19	86.03	96.21	90.19	92.40	87.65
GEAR (**14.0M**)	**75.51**	**90.86**	**81.58**	**96.34**	**90.53**	**96.80**	89.76	**93.80**	**89.40**

Table 2: GLUE performance with DeBERTa-v3-base. We report the mean over three runs with different random seeds. Best results are shown in **bold**.

Method (Params)	CoLA Mcc.	MNLI Acc.	MRPC Acc.	RTE Acc.	QNLI Acc.	SST-2 Acc.	STS-B Corr.	QQP Acc.	Avg.
LoRA (1.9M)	68.31	89.74	90.19	84.16	93.61	95.30	85.20	90.35	87.11
AdaLoRA (1.9M)	70.07	89.93	90.44	88.17	**94.27**	95.13	91.21	**91.41**	88.83
Riemannion (1.9M)	67.41	88.93	89.01	85.13	92.64	94.00	90.12	87.92	86.90
GeoLoRA (1.9M)	68.18	89.67	87.76	87.73	92.93	95.52	90.15	87.77	87.46
FlexLoRA (1.9M)	70.22	90.13	90.67	88.45	94.24	95.34	91.53	90.27	88.85
GEAR (**1.5M**)	**70.85**	**90.97**	**92.46**	**89.53**	**94.27**	**96.22**	**91.83**	90.88	**89.63**

4.3 Natural Language Understanding

We evaluate GEAR with DeBERTa-v3-base (He et al., 2023) on the GLUE benchmark (Wang et al., 2019), covering eight tasks with linguistic objectives, including acceptability, inference, paraphrase detection, semantic similarity, and sentiment classification. The reference rank is set to 8, while GEAR initializes at $r_0 = 4$ with $r_{\max} = 16$ as an upper bound; the final rank is determined by the accepted expansions rather than the limit. We report the task-specific metrics and their average.

As shown in Table 2, GEAR achieves the best average score of 89.63 with only 1.5M trainable parameters, improving over FlexLoRA by 0.78 points while using about 21% fewer parameters. GEAR attains the strongest result on seven of the eight tasks, with particularly clear gains on MRPC, RTE, and STS-B. The improvement across heterogeneous objectives suggests that the benefit of GEAR is not tied to a specific task structure. By expanding the adaptation space according to task-relevant directions beyond its current first-order reach, GEAR can introduce additional capacity more selectively across diverse tasks.

4.4 Visual Task

To assess the generality of GEAR beyond NLP, we evaluate it with ViT-B/16 (Dosovitskiy et al., 2021) on the Visual Task Adaptation Benchmark (VTAB-1K) (Zhai et al., 2020), which covers 19 recognition tasks grouped into Natural, Specialized, and Structured domains. We use the same rank configuration as in the NLU experiments.

As shown in Table 3, GEAR achieves the highest overall average of 73.48 with only 0.2M trainable parameters, compared with 72.55 for FlexLoRA using 0.3M. GEAR obtains the best result on 14 of

Table 3: VTAB-1K performance with ViT-B/16. We report the mean over three runs with different random seeds. Best results are shown in **bold**.

Dataset	LoRA (0.3M)	AdaLoRA (0.3M)	Riemannion (0.3M)	GeoLoRA (0.3M)	FlexLoRA (0.3M)	GEAR (**0.2M**)
Natural						
CIFAR100	69.39	71.81	70.78	68.05	71.30	**71.88**
Caltech101	90.60	89.96	91.03	88.87	91.72	**91.78**
DTD	71.01	**71.76**	71.28	68.88	70.18	71.60
Flowers	98.80	99.04	99.09	99.06	98.98	**99.12**
Pets	89.89	91.58	91.85	91.71	91.14	**91.93**
SVHN	90.53	90.59	89.08	91.39	90.25	**91.89**
SUN397	53.78	**55.61**	55.12	54.63	54.14	55.03
Specialized						
Camelyon	83.26	81.92	81.60	82.16	**83.98**	83.93
EuroSAT	93.74	94.04	93.13	94.61	93.91	**94.91**
RESISC45	85.40	85.37	84.30	84.52	85.41	**85.98**
Retinopathy	73.60	75.40	75.16	**75.41**	74.13	74.91
Structured						
CLEVR-C	80.91	81.23	78.95	80.63	81.30	**82.07**
CLEVR-D	64.74	65.38	68.49	63.37	63.56	**68.51**
DMLab	50.04	50.23	49.95	50.08	52.02	**52.18**
KITTI-D	77.87	78.34	75.81	78.14	**78.95**	78.90
dSpr-Loc	75.76	73.38	72.40	73.62	75.49	**76.78**
dSpr-Ori	47.65	49.32	48.71	48.27	49.02	**49.52**
sNORB-Azi	29.49	28.00	27.76	30.25	29.84	**31.32**
sNORB-Ele	41.93	37.56	39.58	40.99	43.17	**43.89**
Average	72.02	72.13	71.79	71.82	72.55	**73.48**

the 19 tasks, including 7 of the 8 Structured tasks. The gains across visually and semantically diverse domains indicate that the proposed rank-evolution mechanism is not tied to a particular modality or task family. By adapting the low-rank space according to task-relevant directions beyond its current first-order reach, GEAR maintains strong transfer performance while introducing fewer additional degrees of freedom.

Across all evaluated settings, GEAR achieves the strongest average performance with fewer trainable parameters than the strongest adaptive-rank baseline. These gains span two language-model backbones, eight GLUE tasks, and 19 VTAB-1K tasks, indicating robustness across backbones, task families, and modalities. Importantly, they arise from adaptively selected final ranks rather than simply increasing the rank budget, supporting the hypothesis that identifying valuable directions outside the current adaptation space enables more efficient rank growth.

5 Conclusion

We presented GEAR, a geometry-guided framework for dynamic low-rank adaptation that shifts the focus from capacity allocation to the sufficiency of the current adaptation space. We established a first-order information boundary for standard factor gradients, showing that task-relevant normal information is unobservable from the current fixed-rank parameterization, and further related the normal spectrum to the locally optimal value of rank expansion. Building on this analysis, GEAR combines fixed-rank manifold optimization, efficient normal-spectrum estimation, and task-induced geometric initialization to adapt rank according to task-relevant directions beyond the current first-order reachable space. Across commonsense reasoning, GLUE, and VTAB-1K, GEAR achieves the highest average performance in all evaluated settings while using substantially fewer trainable parameters than strong adaptive-rank baselines. Overall, our results suggest that dynamic rank is most effectively viewed not as a budget-allocation problem alone, but as the controlled evolution of the adaptation space in response to previously unrepresented task information.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, LaTeX formatting, and the examination of mathematical derivations. All AI-assisted suggestions were independently reviewed and verified by the authors, who take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We provide complete proofs of all theoretical results and detailed algorithm descriptions in the appendix. Experimental configurations, hyperparameters, dataset protocols, baseline settings, and additional ablations are also provided to facilitate reproducibility. All reported results are averaged over three random seeds.

REFERENCES

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second PASCAL recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, pp. 1–9, 2006.

Luisa Bentivogli, Bernardo Magnini, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. The fifth PASCAL recognizing textual entailment challenge. In *Text Analysis Conference*, 2009.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7432–7439, 2020. doi: 10.1609/aaai.v34i05.6239.

Vladimir Bogachev, Vladimir Aletov, Alexander Molozhavenko, Denis Bobkov, Vera Soboleva, Aibek Alanov, and Maxim Rakhuba. LoRA meets Riemannion: Muon optimizer for parametrization-independent low-rank adapters. In *International Conference on Learning Representations*, 2026.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pp. 1–14. Association for Computational Linguistics, 2017. doi: 10.18653/v1/S17-2001.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2924–2936. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1300.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge, 2018.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, volume 3944 of *Lecture Notes in Computer Science*, pp. 177–190. Springer, 2006. doi: 10.1007/11736790_9.

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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4133–4145. Association for Computational Linguistics, 2023a. doi: 10.18653/v1/2023.emnlp-main.252.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5:220–235, 2023b. doi: 10.1038/s42256-023-00626-4.

William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing*, 2005.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Abdessalam Ed-dib, Zhanibek Datbayev, and Amine M. Aboussalah. GeLoRA: Geometric adaptive ranks for efficient LoRA fine-tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 25174–25196. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1372.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 1–9. Association for Computational Linguistics, 2007.

Aaron Grattafiori et al. The llama 3 herd of models, 2024.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011. doi: 10.1137/090771806.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sE7-XhLxHA.

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.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Lee. LLM-adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5254–5276. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.319.

Muqing Liu, Chongjie Si, and Yuheng Jia. FlexLoRA: Entropy-guided flexible low-rank adaptation. In *International Conference on Learning Representations*, 2026.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32100–32121. PMLR, 2024.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1260.

JuneYoung Park, Minjae Kang, Seongbae Lee, Haegang Lee, Seongwan Kim, and Jaeho Lee. Riemannian optimization for LoRA on the stiefel manifold. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20971–20985, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1143.

Qwen, An Yang, et al. Qwen2.5 technical report, 2025.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1264.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8732–8740, 2020. doi: 10.1609/aaai.v34i05.6399.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 4463–4473. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1454.

Steffen Schotthöfer, Emanuele Zangrando, Gianluca Ceruti, Francesco Tudisco, and Jonas Kusch. GeoLoRA: Geometric integration for parameter efficient fine-tuning. In *International Conference on Learning Representations*, 2025.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642. Association for Computational Linguistics, 2013.

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 *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3274–3287. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.eacl-main.239.

Bart Vandereycken. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013. doi: 10.1137/110845768.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019.

Shaowen Wang, Linxi Yu, and Jian Li. LoRA-GA: Low-rank adaptation with gradient approximation. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641, 2019. doi: 10.1162/tacl_a_00290.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-1101.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1472.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark, 2020.

Fangzhao Zhang and Mert Pilanci. Riemannian preconditioned LoRA for fine-tuning foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59641–59669. PMLR, 2024.

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.

Yuan Zhang, Jiang Hu, Zhijian Lai, Lin Lin, and Zaiwen Wen. Retraction-free optimization over the stiefel manifold for the LoRA fine-tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. doi: 10.1109/TPAMI.2026.3711595.

A Supplementary Overview

This supplementary material provides additional methodological, theoretical, and experimental details for GEAR. Appendix B describes the implementation of fixed-rank retraction and the validation procedure for rank-expansion candidates. Appendix C provides complete proofs of the three theoretical results presented in the main paper. Appendix D reports the datasets, evaluation protocols, training configurations, and hyperparameter settings used in our experiments. Finally, ablation studies are provided in Appendix E.

B Additional Method Details

B.1 Low-Rank Retraction

We detail the retraction used in Section 3.2. Recall that the current adaptation state is represented as $X = UC$, with $U^\top U = I_r$, and let

$$D = U^\top G, \qquad K = (I - UU^\top)GV.$$

Using the tangent-gradient decomposition in Eq. (1), the update before retraction can be written as

$$X^- = U(C - \eta D) - \eta K V^\top = LR,$$

where

$$L = [\,U,\ K\,], \qquad R = \begin{bmatrix} C - \eta D \\ -\eta V^\top \end{bmatrix}.$$

Hence, X^- has rank at most $2r$ and can be manipulated without forming the full $n \times m$ matrix.

We compute thin QR factorizations of the two compact factors,

$$L = Q_L R_L, \qquad R^\top = Q_R R_R,$$

which gives

$$X^- = Q_L \left(R_L R_R^\top\right) Q_R^\top.$$

The matrix $R_L R_R^\top$ is at most $2r \times 2r$. We therefore compute only its compact SVD,

$$R_L R_R^\top = P \Sigma Q^\top,$$

and retain the leading r singular components. The retracted state is

$$X_{t+1} = (Q_L P_r)\Sigma_r (Q_R Q_r)^\top.$$

Accordingly, the updated low-rank representation can be obtained directly as

$$U_{t+1} = Q_L P_r, \qquad V_{t+1} = Q_R Q_r, \qquad C_{t+1} = \Sigma_r V_{t+1}^\top.$$

Thus, the retraction simultaneously restores the rank-r constraint and updates the left and right subspace bases required by the next iteration, while avoiding any SVD of the full adaptation matrix.

B.2 Rank-Expansion Candidate Validation

At each rank audit, the selected normal directions are treated as candidate expansions rather than committed immediately. We order the proposed directions by the utility score in Eq. (5) and construct several prefixes with different birth-step scales. The current adapter and all candidate variants are evaluated on the same set of separate training microbatches under matched stochasticity, so that their losses are directly comparable.

Let $\bar{L}_0$ and $\bar{L}_{\mathrm{cand}}$ denote the mean training losses of the current adapter and a candidate, respectively. We define the relative improvement as

$$\mathrm{RI} = \frac{\bar{L}_0 - \bar{L}_{\mathrm{cand}}}{|\bar{L}_0|}.$$

A candidate is accepted only if $\mathrm{RI} > \rho$, where ρ is the prescribed acceptance threshold.

Among the accepted candidates, we favor the smallest rank increase whose loss is close to the best observed value, and then the smaller birth-step scale. The selected expansion is committed only if it also satisfies the per-module rank limit and the global budget. If no candidate passes the acceptance criterion, the audit leaves the adapter unchanged.

Table 4: Training time and peak CUDA memory on DTD with ViT-B/16.

Method	Training time (s)	Peak CUDA memory (MiB)
LoRA	789.83	3825.15
FlexLoRA	1006.75	6553.87
GEAR	1060.83	4458.21

B.3 Computational and Memory Complexity

GEAR preserves the low-rank computational structure throughout training. For a rank-r adapter, a standard LoRA factor update requires $\mathcal{O}((n+m)r)$ element-wise operations, while the additional retraction in GEAR costs $\mathcal{O}((n+m)r^2+r^3)$ due to thin QR factorizations and an SVD of a $2r \times 2r$ core matrix. Since $r \ll \min\{n,m\}$, this overhead remains confined to the low-rank factors and small core matrices, without introducing any full $n \times m$ decomposition. The committed adapter state requires only $\mathcal{O}((n+m)r)$ memory.

TGI is performed once before fine-tuning. In the language experiments, its $2(p+1)$ probing stages require $2(p+1)$ calibration backward passes per initialization run and $\mathcal{O}(h\max\{n,m\})$ peak sketch storage per adapted matrix. The QR and SVD operations cost $\mathcal{O}((p+1)(n+m)h^2)$. We use $p=1$ and $h \ll \min\{n,m\}$. In VTAB, initialization instead aggregates dense gradient proxies for the selected Q/V projections from three calibration batches, requiring $\mathcal{O}(nm)$ temporary gradient storage per adapted matrix.

Rank audits introduce additional computation only at scheduled audit steps. With the same probe width h, the projections, orthogonalization, and decomposition of the compressed normal sketch incur $\mathcal{O}((n+m)rh+(n+m)h^2)$ low-rank linear-algebra cost and $\mathcal{O}((n+m)h)$ temporary memory, in addition to the backward probes used to obtain gradient products. Both r and h are small relative to the ambient dimensions, and no full gradient or normal matrix is explicitly stored.

The main audit overhead therefore comes from the additional backward probes and candidate forward evaluations rather than large-scale matrix operations. These computations are performed only for $t \in \mathcal{T}_{\text{audit}}$, while ordinary training steps incur only the fixed-rank manifold update. Moreover, all probing variables are transient and do not contribute to the committed adapter rank or the trainable-parameter counts reported in our experiments.

We further assess the measured resource costs on the DTD task with ViT-B/16 in Table 4.

C Proofs

C.1 Proof of Theorem 1: Factor-Gradient Observability Barrier

Proof. Consider the factorization map

$$\Phi : \mathbb{R}^{n\times r} \times \mathbb{R}^{r\times m} \to \mathbb{R}^{n\times m}, \qquad \Phi(B,A) = BA.$$

Its differential at (B,A) is

$$D\Phi_{(B,A)}[\Delta B, \Delta A] = \Delta B A + B \Delta A.$$

For arbitrary perturbations $(\Delta B, \Delta A)$, define $B(t) = B + t\Delta B$ and $A(t) = A + t\Delta A$. Since the set of full-rank matrices is open, both factors remain full rank for sufficiently small t. Hence $X(t) = B(t)A(t) \in \mathcal{M}_r$ locally around $t=0$, and

$$\dot{X}(0) = \Delta B A + B \Delta A = D\Phi_{(B,A)}[\Delta B, \Delta A].$$

Therefore,

$$\operatorname{Im} D\Phi_{(B,A)} \subseteq T_X \mathcal{M}_r.$$

To determine the dimension of this image, consider the kernel condition $\Delta B A + B \Delta A = 0$. Since A has full row rank, it admits a right inverse A_R satisfying $AA_R = I_r$. Right-multiplication by A_R

gives $\Delta B = -B\Delta A A_R$, so $\Delta B = BK$ for some $K \in \mathbb{R}^{r\times r}$. Substitution into the kernel condition yields $B(KA + \Delta A) = 0$. Since B has full column rank, $\Delta A = -KA$. Conversely, every pair $(BK, -KA)$ satisfies $D\Phi_{(B,A)}[BK, -KA] = 0$. Thus,

$$\ker D\Phi_{(B,A)} = \{(BK, -KA) : K \in \mathbb{R}^{r\times r}\}.$$

The parametrization by K is injective because B has full column rank; hence $\dim \ker D\Phi_{(B,A)} = r^2$. By the rank–nullity theorem,

$$\dim \operatorname{Im} D\Phi_{(B,A)} = nr + rm - r^2 = r(n + m - r).$$

Since $\dim T_X\mathcal{M}_r = \dim \mathcal{M}_r = r(n + m - r)$, the inclusion above is between spaces of equal dimension, and therefore

$$\operatorname{Im} D\Phi_{(B,A)} = T_X\mathcal{M}_r.$$

We next identify the adjoint of $D\Phi_{(B,A)}$. Under the Frobenius inner product,

$$\begin{aligned}\left\langle G, D\Phi_{(B,A)}[\Delta B, \Delta A]\right\rangle_F &= \langle G, \Delta BA + B\Delta A\rangle_F \\ &= \langle GA^\top, \Delta B\rangle_F + \langle B^\top G, \Delta A\rangle_F.\end{aligned}$$

Hence

$$D\Phi^*_{(B,A)}(G) = (GA^\top, B^\top G) = \mathcal{O}_{A,B}(G).$$

For any linear operator L between finite-dimensional inner-product spaces, $\ker L^* = (\operatorname{Im} L)^\perp$. Applying this identity to $D\Phi_{(B,A)}$ and using the characterization above gives

$$\ker \mathcal{O}_{A,B} = \left(\operatorname{Im} D\Phi_{(B,A)}\right)^\perp = (T_X\mathcal{M}_r)^\perp = N_X\mathcal{M}_r.$$

Finally, for $D = G_1 - G_2$, $\mathcal{O}_{A,B}(G_1) = \mathcal{O}_{A,B}(G_2)$ if and only if $\mathcal{O}_{A,B}(D) = 0$, which is equivalent to $D \in N_X\mathcal{M}_r$. Equivalently, the tangent projection of D vanishes, and hence

$$(G_1)_T = (G_2)_T.$$

This proves both statements. □

C.2 Proof of Theorem 2: Local Optimality of Rank Expansion

Proof. Let $X = U\Sigma V^\top \in \mathcal{M}_r$ be a compact SVD, and let $U_\perp$ and $V_\perp$ be orthonormal complements of U and V. For any perturbation H, write

$$\begin{bmatrix} U^\top \\ U_\perp^\top \end{bmatrix} H \begin{bmatrix} V & V_\perp \end{bmatrix} = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}.$$

Its normal component is $H_N = U_\perp H_{22} V_\perp^\top$, while the remaining three blocks constitute the tangent component H_T.

We first characterize the tangent cone of $\mathcal{V}_{r+k} = \{Y : \operatorname{rank}(Y) \le r + k\}$ at X. In the above coordinates,

$$\mathcal{C}_{r,k}(X) = \{H_T + H_N : H_T \in T_X\mathcal{M}_r,\ H_N \in N_X\mathcal{M}_r,\ \operatorname{rank}(H_N) \le k\}.$$

To verify this characterization, consider a feasible curve $X(t) \in \mathcal{V}_{r+k}$ with $X(0) = X$ and $\dot X(0) = H$. Its block form is

$$X(t) = \begin{bmatrix} \Sigma + tH_{11} + o(t) & tH_{12} + o(t) \\ tH_{21} + o(t) & tH_{22} + o(t) \end{bmatrix}.$$

For sufficiently small t, the upper-left block is nonsingular. Since $\operatorname{rank}(X(t)) \le r + k$, its Schur complement has rank at most k. Dividing by t and letting $t \to 0$ therefore gives $\operatorname{rank}(H_{22}) \le k$.

Conversely, suppose $\operatorname{rank}(H_{22}) \le k$ and define $A_t = \Sigma + tH_{11}$. Consider

$$\widetilde X(t) = \begin{bmatrix} A_t & tH_{12} \\ tH_{21} & tH_{22} + t^2 H_{21} A_t^{-1} H_{12} \end{bmatrix}.$$

Its Schur complement with respect to A_t is tH_{22}, so $\operatorname{rank}(\widetilde{X}(t)) = r + \operatorname{rank}(H_{22}) \le r + k$. Moreover, $\widetilde{X}(0) = X$ and $\dot{\widetilde{X}}(0) = H$, establishing the stated tangent-cone characterization.

We now minimize

$$q_\eta(H) = \langle G, H\rangle_F + \frac{1}{2\eta}\|H\|_F^2$$

over $\mathcal{C}_{r,k}(X)$. Since the tangent and normal spaces are orthogonal, write $G = G_T + G_N$ and $H = H_T + H_N$. Then

$$q_\eta(H) = \left(\langle G_T, H_T\rangle_F + \frac{1}{2\eta}\|H_T\|_F^2\right) + \left(\langle G_N, H_N\rangle_F + \frac{1}{2\eta}\|H_N\|_F^2\right),$$

subject only to $\operatorname{rank}(H_N) \le k$.

For the tangent component,

$$\langle G_T, H_T\rangle_F + \frac{1}{2\eta}\|H_T\|_F^2 = \frac{1}{2\eta}\|H_T + \eta G_T\|_F^2 - \frac{\eta}{2}\|G_T\|_F^2,$$

so the unique minimizer is $H_T^\star = -\eta G_T$.

Similarly, for the normal component,

$$\langle G_N, H_N\rangle_F + \frac{1}{2\eta}\|H_N\|_F^2 = \frac{1}{2\eta}\|H_N + \eta G_N\|_F^2 - \frac{\eta}{2}\|G_N\|_F^2.$$

Thus, under $\operatorname{rank}(H_N) \le k$, minimizing the normal term is equivalent to finding the best rank-k approximation of $-\eta G_N$ in Frobenius norm. By the Eckart–Young–Mirsky theorem, $H_N^\star = -\eta(G_N)_k$. Hence

$$H_k^\star = -\eta\left[G_T + (G_N)_k\right].$$

It remains to evaluate the additional decrease relative to fixed-rank optimization. Over $T_X\mathcal{M}_r$, $\min q_\eta = -\frac{\eta}{2}\|G_T\|_F^2$, whereas over $\mathcal{C}_{r,k}(X)$,

$$\min q_\eta = -\frac{\eta}{2}\|G_T\|_F^2 - \frac{\eta}{2}\|(G_N)_k\|_F^2.$$

Therefore,

$$\Delta_k = \frac{\eta}{2}\|(G_N)_k\|_F^2 = \frac{\eta}{2}\sum_{i=1}^{k}\sigma_i^2(G_N).$$

Thus, the leading normal singular directions yield the optimal local rank-k expansion, and their squared singular values determine the additional decrease of the local quadratic model. □

C.3 Proof of Theorem 3: Stability under Inexact Normal Estimation

Proof. For brevity, let $N = G_N$, $\widehat{N} = \widehat{G}_N$, $N_k = (G_N)_k$, and $\widehat{N}_k = (\widehat{G}_N)_k$. By assumption, $\|\widehat{N} - N\|_F \le \varepsilon$ and $\tau_k = \|N - N_k\|_F$.

Since N_k and $\widehat{N}_k$ lie in the normal space, they are orthogonal to the tangent component of G. Hence, for any normal matrix B,

$$q_\eta(-\eta B) = \frac{\eta}{2}\|N - B\|_F^2 - \frac{\eta}{2}\|N\|_F^2.$$

Applying this identity to $B = N_k$ and $B = \widehat{N}_k$ yields

$$q_\eta(-\eta\widehat{N}_k) - q_\eta(-\eta N_k) = \frac{\eta}{2}\left(\|N - \widehat{N}_k\|_F^2 - \|N - N_k\|_F^2\right).$$

Because N_k is a best rank-k approximation of N and $\operatorname{rank}(\widehat{N}_k) \le k$, $\|N - N_k\|_F \le \|N - \widehat{N}_k\|_F$. Therefore, $q_\eta(-\eta\widehat{N}_k) - q_\eta(-\eta N_k) \ge 0$.

It remains to establish the upper bound. Since $\widehat{N}_k$ is a best rank-k approximation of $\widehat{N}$, $\|\widehat{N}-\widehat{N}_k\|_F \leq \|\widehat{N} - N_k\|_F$. Using the triangle inequality,

$$\begin{aligned}\|N - \widehat{N}_k\|_F &\leq \|N - \widehat{N}\|_F + \|\widehat{N} - \widehat{N}_k\|_F \\ &\leq \varepsilon + \|\widehat{N} - N_k\|_F \\ &\leq \varepsilon + \|\widehat{N} - N\|_F + \|N - N_k\|_F \\ &\leq \tau_k + 2\varepsilon.\end{aligned}$$

Substituting this bound into the preceding identity gives

$$\begin{aligned}q_\eta(-\eta\widehat{N}_k) - q_\eta(-\eta N_k) &\leq \frac{\eta}{2}\left[(\tau_k + 2\varepsilon)^2 - \tau_k^2\right] \\ &= 2\eta\varepsilon(\tau_k + \varepsilon).\end{aligned}$$

Consequently,

$$0 \leq q_\eta\Big(-\eta(\widehat{G}_N)_k\Big) - q_\eta(-\eta(G_N)_k) \leq 2\eta\varepsilon(\tau_k + \varepsilon).$$

This completes the proof. □

D Experimental Details

D.1 Details on Commonsense Reasoning Task

The commonsense reasoning suite comprises eight benchmarks: BoolQ (Clark et al., 2019), PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), HellaSwag (Zellers et al., 2019), WinoGrande (Sakaguchi et al., 2020), ARC-e/ARC-c (Clark et al., 2018), and OBQA (Mihaylov et al., 2018). Following the standard multi-task protocol, we combine the training portions of these tasks into the Commonsense170K corpus (Hu et al., 2023) and evaluate each benchmark separately on its corresponding test set. We use LLaMA3-8B and Qwen2.5-7B as backbone models. The main hyperparameter settings are summarized in Table 5.

The probing configuration estimates task-induced gradient structure for geometric initialization. We use probe width $h = 24$ for both TGI and rank audits. For rank audits, each calibration batch contains 128 examples, with 16 from each task. The damping coefficient $\lambda = 0.1$ regularizes the gradient-scale normalization. Subsequent experiments follow the same initialization and rank-audit framework with task-specific adjustments; the VTAB implementation is detailed in Appendix D.3.

D.2 Details on Natural Language Understanding Task

For the natural language understanding experiments, we fine-tune DeBERTa-v3-base on the GLUE benchmark (Wang et al., 2019), comprising eight tasks across single-sentence classification, semantic similarity and paraphrase detection, and natural language inference. Specifically, we evaluate on CoLA (Warstadt et al., 2019), SST-2 (Socher et al., 2013), MRPC (Dolan & Brockett, 2005), QQP (Wang et al., 2019), STS-B (Cer et al., 2017), MNLI (Williams et al., 2018), QNLI (Rajpurkar et al., 2016), and RTE (Dagan et al., 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009). Dataset statistics and evaluation metrics are summarized in Table 6, and the corresponding training and rank-adaptation settings are reported in Table 7. We use the same probing and subspace estimation procedure as in the commonsense reasoning experiments, with shared probe width $h = 24$ for TGI and rank audits and task-specific calibration data and optimization settings. Separate learning rates are used for the adapter and task head because they operate under different parameterizations and effective update scales: the head follows standard Euclidean optimization, whereas the adapter learning rate controls the manifold-constrained update before retraction.

D.3 Details on Visual Task

We evaluate GEAR with ViT-B/16 on VTAB-1K (Zhai et al., 2020), which contains 19 visual classification tasks across Natural, Specialized, and Structured categories. Dataset statistics and hyperparameter settings are reported in Tables 8 and 9, respectively. For TGI, we aggregate dense gradient proxies for the selected Q/V projections from three calibration batches. Rank audits and candidate validation use task-specific settings.

Table 5: Hyperparameter settings for LLaMA3-8B and Qwen2.5-7B in commonsense reasoning experiments.

Hyperparameter	**Setting**
Initial effective rank r_0	4
Maximum per-module rank $r_{\max}$	32
Weighted-average rank budget $\bar{r}_{\text{budget}}$	16
Target modules	Q, K, V, O, Up, Down, Gate
Probe Ω width h	24
Probe batch size	128 total (16 per task)
Damping λ	0.1
First audit step t_{start}	100
Audit interval Δt	50
Rank-audit cutoff t_{final}	500
Max. rank increment per module per audit	4
Acceptance threshold ρ	0.5%
Learning rate	2×10^{-2}
Dropout	0.05
Scheduler / warmup	Linear / 0.1
Batch configuration	16 microbatch $\times$ 4 accumulation
Epochs	2

Table 6: Details of the GLUE dataset.

Dataset	**Task**	**# Train**	**# Dev**	**# Test**	**# Label**	**Metrics**
Single-Sentence Classification						
CoLA	Acceptability	8.5k	1k	1k	2	Matthews corr.
SST-2	Sentiment	67k	872	1.8k	2	Accuracy
Similarity and Paraphrase						
MRPC	Paraphrase	3.7k	408	1.7k	2	Accuracy
QQP	Paraphrase	364k	40k	391k	2	Accuracy
STS-B	Similarity	7k	1.5k	1.4k	1	Pearson/Spearman corr.
Natural Language Inference						
MNLI	NLI	393k	20k	20k	3	Accuracy
QNLI	QA/NLI	108k	5.7k	5.7k	2	Accuracy
RTE	NLI	2.5k	276	3k	2	Accuracy

Table 7: Hyper-parameter settings of our method on the GLUE benchmark.

Dataset	**Adapter LR**	**Head LR**	**Batch**	**Epochs**	r_0	$\bar{r}_{\text{budget}}$	$r_{\max}$	t_{start}	Δt	t_{final}
MNLI	5×10^{-4}	5×10^{-4}	32	6	4	8	16	$0.1T$	$0.075T$	$0.7T$
RTE	3.6×10^{-3}	1.2×10^{-3}	32	50	4	8	16	$0.1T$	$0.075T$	$0.7T$
QNLI	3×10^{-3}	3×10^{-4}	32	5	4	8	16	$0.1T$	$0.075T$	$0.7T$
MRPC	3×10^{-3}	1×10^{-3}	32	30	4	8	16	$0.1T$	$0.075T$	$0.7T$
QQP	1.6×10^{-3}	8×10^{-4}	32	5	4	8	16	$0.1T$	$0.075T$	$0.7T$
SST-2	5×10^{-3}	3×10^{-4}	32	24	4	8	16	$0.1T$	$0.075T$	$0.7T$
CoLA	1.6×10^{-3}	8×10^{-4}	32	20	4	8	16	$0.1T$	$0.075T$	$0.7T$
STS-B	4.4×10^{-3}	2.2×10^{-3}	32	25	4	8	16	$0.1T$	$0.075T$	$0.7T$

T denotes the total number of optimization steps. At each rank audit, each module may add at most one new rank-one direction, with an acceptance threshold of $\rho = 0.1\%$ relative loss improvement.

Table 8: Details of the VTAB-1K benchmark.

Dataset	# Classes	# Train	# Val	# Test
Natural				
CIFAR100	100	800/1000	200	10,000
Caltech101	102	800/1000	200	6,084
DTD	47	800/1000	200	1,880
Flowers102	102	800/1000	200	6,149
Pets	37	800/1000	200	3,669
SVHN	10	800/1000	200	26,032
SUN397	397	800/1000	200	21,750
Specialized				
Camelyon	2	800/1000	200	32,768
EuroSAT	10	800/1000	200	5,400
RESISC45	45	800/1000	200	6,300
Retinopathy	5	800/1000	200	42,670
Structured				
CLEVR-Count	8	800/1000	200	15,000
CLEVR-Dist	6	800/1000	200	15,000
DMLab	6	800/1000	200	22,735
KITTI-Dist	4	800/1000	200	711
dSpr-Loc	16	800/1000	200	73,728
dSpr-Ori	16	800/1000	200	73,728
sNORB-Azim	18	800/1000	200	12,150
sNORB-Ele	18	800/1000	200	12,150

Table 9: Hyperparameter settings for the VTAB-1K experiments.

Hyperparameter	Setting
Initial effective rank r_0	4
Weighted-average rank budget $\bar{r}_{\text{budget}}$	8
Maximum per-module rank $r_{\max}$	16
Learning rate	5×10^{-3}
First audit step t_{start}	160
Audit interval Δt	50
Rank-audit cutoff t_{final}	1300
Acceptance threshold ρ	0.03%
Batch size	64
Epochs	100

E Ablation Studies

To isolate the contribution of each component, we conduct a component-wise ablation on LLaMA3-8B over the commonsense reasoning suite. Starting from the standard LoRA baseline, we examine task-induced geometric initialization (TGI), fixed-rank manifold optimization, and geometry-guided rank expansion. The results are reported in Table 10.

TGI and manifold optimization independently improve the average score from 85.03 to 85.57 and 85.65, respectively, while combining them further raises it to 86.23. This suggests that task-informed initialization and geometry-preserving optimization provide complementary benefits in constructing and optimizing the current adaptation space.

Rank expansion brings a larger gain. Adding it to manifold optimization improves the average from 85.65 to 87.32, while combining all three components further increases the score to 88.45, with improvements across all eight tasks. TGI also remains beneficial when rank expansion is enabled, improving the manifold-plus-expansion configuration from 87.32 to 88.45. Taken together, these ablations align with the central view of GEAR: dynamic rank adaptation should distinguish between better exploiting the current adaptation space and expanding it when task-relevant directions remain unrepresented.

Table 10: Component ablation of GEAR on commonsense reasoning with LLaMA3-8B. TGI denotes task-induced geometric initialization.

TGI	Manifold	Rank Expansion	BoolQ	PIQA	SIQA	HellaS.	WinoG.	ARC-e	ARC-c	OBQA	Avg.
–	–	–	72.37	87.83	81.34	94.81	85.08	91.46	81.32	86.00	85.03
✓	–	–	72.51	87.57	81.27	94.88	86.94	92.16	83.08	86.16	85.57
–	✓	–	72.40	87.03	81.12	94.46	87.38	92.51	83.12	87.21	85.65
✓	✓	–	73.15	89.13	82.01	94.90	87.27	92.88	83.30	87.17	86.23
–	✓	✓	73.95	89.73	83.13	95.84	88.66	93.04	84.71	89.51	87.32
✓	✓	✓	**75.19**	**90.81**	**83.78**	**96.82**	**89.82**	**93.77**	**85.58**	**91.80**	**88.45**