AR1-ZO: Topology-Aware Rank-1 Zeroth-Order Queries for High-Rank LoRA Fine-Tuning

Ziye Chen[1] **Hongbin Lin**[1] **Chenyu Zhang**[1]
Xiangda Yan[2] **Yongjie Yang**[2] **Yao Shu**[1,3]*

[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Xiaomi Inc.
[3]The Hong Kong University of Science and Technology

Abstract

Zeroth-order (ZO) optimization enables large-language-model fine-tuning without storing backpropagation activations, while LoRA supplies compact trainable adapters. Combining them creates a rank paradox: increasing LoRA rank improves adapter capacity, but standard two-point ZO either perturbs a rank-dependent number of coordinates or, under atomwise updates, can make the finite-difference signal unobservable. This paper shows that the bottleneck is a measurement-topology problem rather than a need for an external subspace. LoRA already decomposes into matched rank-1 atoms, each a complete factor-coordinate block of dimension $d_{\text{out}} + d_{\text{in}}$. Querying one atom per step keeps the stored adapter rank r while removing r from the single-query perturbation dimension. The naive atomwise query is still miscalibrated: if it inherits canonical LoRA scaling α/r, the active finite-difference signal shrinks as $1/r$ and the active finite-difference signal-to-noise ratio (FD-SNR) as $1/r^2$, producing directional collapse under a fixed residual evaluation-noise floor. AR1-ZO pairs alternating rank-1 atom queries with topology-aware scaling $\gamma = \alpha r$, restoring rank-invariant active signal without auxiliary bases, activation hooks, curvature estimates, or extra forward queries. Theory proves atom minimality, rank-independent active query dimension, directional collapse and restoration, and the remaining rank dependence as an amortized coverage cost. Experiments on OPT and Qwen3 models validate the signal mechanism and show that AR1-ZO makes high-rank LoRA effective among matched-budget ZO methods under the standard two-forward-pass query budget.

1 Introduction

Fine-tuning large language models (LLMs) under tight memory budgets requires controlling both activation storage and trainable degrees of freedom. Zeroth-order (ZO) optimization, exemplified by MeZO [14], removes the activation-memory cost of backpropagation by estimating descent directions from forward-pass scalar differences. LoRA [8], together with related low-rank adaptation variants such as QLoRA [4], DoRA [13], and AdaLoRA [26], reduces the number of trainable coordinates by freezing pretrained weights and learning compact low-rank updates. Their combination is therefore natural, but it turns the LoRA rank into a double-edged quantity: the same rank that gives the adapter capacity can also determine what a two-point ZO oracle has to perturb and whether its scalar loss difference remains visible.

*Correspondence to Yao Shu <yaoshu@hkust-gz.edu.cn>.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

The usual explanation is the curse of dimensionality. Classical randomized gradient estimators incur variance that grows with the perturbation dimension [15, 12]; if a ZO query perturbs all LoRA factors at once, the dimension grows with rank. Existing remedies reduce the effective dimension through block-coordinate perturbations [16], random subspaces [25], low-rank estimators [3, 19, 21], LoRA-manifold tangent updates [20], activation-guided perturbations [10], and second-order reweighting [27]. Recent low-rank and geometry-aware methods such as TeZO and RoZO further show that temporal low-rank structure and LoRA-manifold geometry can improve ZO fine-tuning, but they still leave open the atom-level measurement unit and scale of a two-forward-pass LoRA-ZO oracle. In parallel, LoRA scaling and optimization studies such as rsLoRA [9], RoRA [11], and LoRA+ [7] refine rank normalization or factor-wise dynamics under first-order joint updates. These lines establish that both query geometry and scaling matter, but they do not settle the measurement unit of black-box LoRA-ZO: **Can the rank-1 atoms already inside LoRA be used as complete two-forward-pass ZO query blocks, and what scaling keeps those atomwise measurements observable at high rank?** A fuller discussion of the related literature and novelty boundary appears in Appendix A.

Our answer starts by separating three roles that rank often conflates: stored adapter capacity, single-query perturbation dimension, and finite-difference signal scale. The LoRA product can be written as a sum of rank-1 atoms, but the contribution is not this standard algebraic identity. The key object is the matched column-row pair $(\boldsymbol{b}_k, \boldsymbol{a}_k)$, which is the finest complete LoRA-native factor block that can be exposed to a two-point oracle: perturbing only one side is incomplete, while grouping multiple atoms reintroduces a larger query dimension. An atomwise query therefore keeps the adapter rank r in storage and in long-run coverage, but changes the instantaneous ZO problem from a rank-r factor perturbation to a single complete atom perturbation. This removes the first bottleneck without invoking an auxiliary basis, activation hook, curvature estimate, or external subspace.

The same native query topology also reveals why the obvious atomwise optimizer fails. Canonical LoRA scaling uses the coefficient α/r, calibrated for joint updates in which all r atoms contribute together. Under a sequential rank-1 ZO measurement, only one atom differs between the positive and negative endpoints, so inheriting the joint coefficient suppresses the active finite-difference numerator by $1/r$. Residual evaluation noise is not reduced by the same factor, making the active FD-SNR decay as $1/r^2$ and eventually erasing directional fidelity before any learning-rate multiplier can act. We call this failure **directional collapse** induced by a **topology-scaling mismatch**. AR1-ZO resolves it by assigning the queried atom a rank-invariant active coefficient through the topology-aware calibration $\gamma = \alpha r$. This is not a new joint LoRA scaling rule; it is a measurement calibration for the sequential two-point oracle.

The resulting optimizer, Alternating Rank-1 Zeroth-Order optimization (AR1-ZO), cycles through complete rank-1 atoms while using the corrected active measurement scale. Each step still uses the standard two forward evaluations, and in transformer implementations the same active rank index can be shared across adapted matrices. The theoretical accounting then becomes precise: rank no longer enters the single-query perturbation dimension or the active FD-SNR; it remains as the amortized cost of revisiting all stored atoms. This yields three contributions.

- **High rank need not be a single-query ZO dimension tax.** We formalize the matched atom $(\boldsymbol{b}_k, \boldsymbol{a}_k)$ as the minimal complete LoRA-native query block and show that its active two-point estimator has perturbation dimension $d_{\text{out}} + d_{\text{in}}$, not $r(d_{\text{out}} + d_{\text{in}})$, while preserving a rank-r adapter.

- **The high-rank failure of naive atomwise ZO is a measurement collapse.** We prove that inheriting the joint-update scale α/r makes active FD-SNR decay as $1/r^2$, derive the topology-aware correction $\gamma = \alpha r$ that restores rank-invariant active signal, and extend the accounting to rank-m blocks and other joint-normalized factorizations.

- **AR1-ZO makes the accounting operational under a pure black-box budget.** We instantiate alternating atom queries with two forward passes per step, prove the remaining rank dependence is an amortized coverage cost and characterize when aligned native atoms carry dense descent signal, then validate the mechanism with spectral-alignment diagnostics, rank-dependent active FD-SNR sweeps, downstream tasks, scaling-law controls, and efficiency measurements.

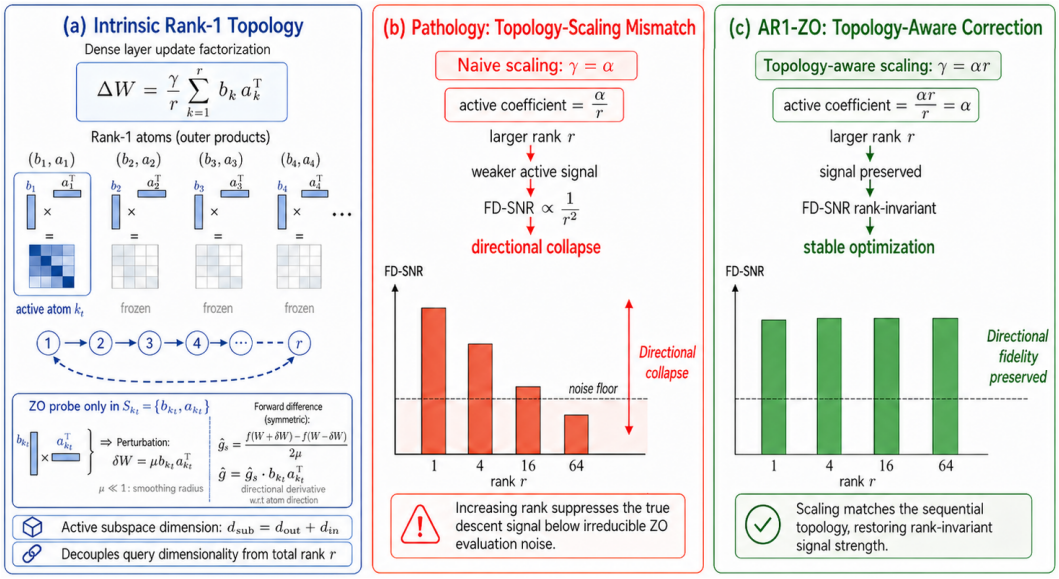


Figure 1: The Optimization Pathology of LoRA-ZO and the AR1-ZO Solution.

2 Preliminaries: LoRA and Zeroth-Order Optimization

Applying ZO to LoRA is governed by three separable quantities: the trainable low-rank coordinates and output coefficient, the coordinates moved by one two-point query, and the evaluation noise that determines whether the scalar loss difference is observable. A black-box LoRA optimizer therefore faces two constraints: the queried coordinate set should not grow with rank, and the finite-difference numerator should not be scaled below the noise floor.

LoRA parameterization and rank scaling. LoRA turns dense adaptation into a factored low-rank parameterization. For a frozen dense layer $\boldsymbol{W}_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ and input activation x, LoRA injects a trainable rank-r update through two factors $\boldsymbol{B} \in \mathbb{R}^{d_{\text{out}} \times r}$ and $\boldsymbol{A} \in \mathbb{R}^{r \times d_{\text{in}}}$:

$$h = \boldsymbol{W}_0 x + \Delta \boldsymbol{W} x, \qquad \Delta \boldsymbol{W} = \frac{\gamma}{r} \boldsymbol{B}\boldsymbol{A}. \tag{1}$$

The canonical choice $\gamma = \alpha$ gives coefficient α/r [8], coupling rank to both adapter capacity and the output change induced by factor movement.

Two-point ZO estimator and perturbation dimension. ZO fine-tuning does not observe gradients of the LoRA factors directly; it observes how the loss changes under a random perturbation of the trainable coordinates. In the standard two-point Randomized Gradient Estimator (RGE) [15], a perturbation $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}_D)$ and smoothing radius μ produce

$$\hat{\nabla}\mathcal{L}(\theta) = \frac{\mathcal{L}(\theta + \mu \boldsymbol{z}) - \mathcal{L}(\theta - \mu \boldsymbol{z})}{2\mu} \boldsymbol{z}. \tag{2}$$

Here D is the number of coordinates perturbed by one two-point query. Under standard L-smooth ZO analyses [15, 12], the mean squared error scales as $\mathcal{O}(D\|\nabla\mathcal{L}(\theta)\|^2 + L^2\mu^2 D^3)$. Perturbing all LoRA factors of one matrix sets $D = r(d_{\text{out}} + d_{\text{in}})$, so full-adapter ZO makes the rank chosen for capacity also control the estimator-variance dimension.

Finite-difference signal and evaluation noise. The dimension term explains why large perturbation spaces are difficult, but it does not fully determine whether a ZO query is useful. The estimator is formed from the clean finite-difference numerator

$$\Delta_\mu(\theta; \boldsymbol{z}) = \mathcal{L}(\theta + \mu \boldsymbol{z}) - \mathcal{L}(\theta - \mu \boldsymbol{z}), \tag{3}$$

which carries the directional information in Eq. (2). In stochastic LLM fine-tuning, paired evaluations on the same mini-batch can reduce data-sampling variance, but residual oracle variation from numerical kernels, stochastic components, or controlled noise injection still enters the observed numerator. We write the observed numerator as $\Delta_\mu(\theta; \boldsymbol{z}) + \xi_+ - \xi_-$, where ξ_+ and ξ_- are the residual

errors of the positive and negative evaluations and σ_ξ^2 denotes their variance scale after shared-batch control. A perturbation is informative only if its induced Δ_μ remains resolvable against this noise. Dimension reduction therefore solves only half of the ZO-LoRA problem: a small coordinate block can still be useless if the coefficient γ/r in Eq. (1) suppresses its output effect before the loss difference is measured.

3 AR1-ZO: Rank-1 Atom Queries with Topology-Aware Scaling

These constraints leave a narrow design target. Full-adapter ZO keeps the LoRA parameterization but perturbs all $r(d_{\text{out}}+d_{\text{in}})$ factor coordinates, so the rank chosen for capacity immediately becomes a variance multiplier. Existing dimension-reduction methods remove this tax with auxiliary bases, activation hooks, curvature side information, or manifold-level geometry. AR1-ZO instead changes the query topology of the existing LoRA factors. Its first step is to expose one complete LoRA atom to the two-point oracle, making the per-query coordinate set independent of r while the stored adapter remains rank r. Its second step is to recalibrate the scale of that sequential measurement: once only one atom moves between the two endpoints, the joint LoRA coefficient α/r no longer reflects the amount of signal that the queried block must carry.

3.1 Complete Rank-1 Atoms as Native ZO Query Blocks

Fix one adapted matrix and write $\Delta \boldsymbol{W} = (\gamma/r)\sum_{k=1}^{r} \boldsymbol{b}_k \boldsymbol{a}_k^\top$, where $\boldsymbol{b}_k$ is the kth column of $\boldsymbol{B}$ and $\boldsymbol{a}_k^\top$ is the kth row of $\boldsymbol{A}$. In first-order LoRA this decomposition is internal to joint-factor optimization, but in ZO the exposed coordinates define the search space. The matched atom k is the pair $(\boldsymbol{b}_k, \boldsymbol{a}_k)$: its first-order variations generate all local changes $\delta \boldsymbol{b}_k \boldsymbol{a}_k^\top + \boldsymbol{b}_k \delta \boldsymbol{a}_k^\top$. It is also the finest complete LoRA-native block, since perturbing one side is incomplete and grouping $m > 1$ atoms raises the block dimension to $m(d_{\text{out}} + d_{\text{in}})$.

At iteration t, a schedule selects $k(t) \in \{1, \dots, r\}$, exposes atom $k(t)$ to the oracle, and holds all other atoms fixed during the two endpoint evaluations. Let $q = d_{\text{out}} + d_{\text{in}}$ be the number of stored coordinates in one atom, let $\boldsymbol{z} \in \mathbb{R}^q$ be the active query direction, and let $\boldsymbol{P}_t$ inject $\boldsymbol{z}$ into the selected factor slots while zeroing all other LoRA coordinates. We write $\hat{\nabla}_{k(t)}\mathcal{L}$ for the resulting estimate of the gradient with respect to the active atom coordinates:

$$\hat{\nabla}_{k(t)}\mathcal{L} = \frac{\mathcal{L}(\theta + \mu \boldsymbol{P}_t \boldsymbol{z}) - \mathcal{L}(\theta - \mu \boldsymbol{P}_t \boldsymbol{z})}{2\mu} \boldsymbol{z}, \tag{4}$$

using the same two loss evaluations as standard ZO. Equation (4) changes the dimension term without changing the information model: the oracle still observes two scalar losses, and frozen atoms are shared by both endpoints. The stored adapter remains rank r, but one matrix query perturbs $d_{\text{out}}+d_{\text{in}}$ rather than $r(d_{\text{out}} + d_{\text{in}})$ coordinates. Rank becomes a coverage index rather than an instantaneous variance multiplier. The active numerator, however, still inherits γ/r from Eq. (1).

3.2 Topology-Aware Scaling: $\gamma = \alpha r$

Atomwise querying changes how the jointly normalized adapter is measured. The standard choice $\gamma = \alpha$ assumes a joint update in which all r atoms contribute through the coefficient α/r. Under Eq. (4), however, only one atom differs between the positive and negative endpoints, so that single atom must carry the observable finite-difference signal. To isolate this topology effect, fix the user-level active coefficient α and set

$$\gamma = \alpha r, \qquad \frac{\gamma}{r} = \alpha. \tag{5}$$

With this calibration, the forward map seen by an active-atom query decomposes into an active term with rank-invariant coefficient and a frozen background shared by both endpoints:

$$h_t(x) = \boldsymbol{W}_0 x + \underbrace{\alpha\, \boldsymbol{b}_{k(t)} \boldsymbol{a}_{k(t)}^\top x}_{\text{active atom}} + \underbrace{\alpha \sum\nolimits_{j \neq k(t)} \boldsymbol{b}_j \boldsymbol{a}_j^\top x}_{\text{frozen background}}. \tag{6}$$

Equations (5)–(6) separate query dimension from measured signal scale at the model-output level. The active query dimension has already been reduced to q, and the frozen background can shape

the current point, but it is identical in the two endpoint evaluations and therefore cannot rescue a suppressed active numerator. Under the naive fixed-scale choice $\gamma = \alpha$, the active coefficient would be α/r, so the finite-difference signal amplitude is $\mathcal{O}(\alpha/r)$ and the signal power is $\mathcal{O}(\alpha^2/r^2)$ whenever the residual noise floor is rank-independent. Increasing rank can then increase representational capacity while making each atomwise measurement less reliable. A larger learning rate cannot repair this failure, because it multiplies the noisy estimate only after the finite-difference numerator has already lost directional information. Section 4.2 formalizes this collapse, and Section 4.3 proves restoration under topology-aware scaling.

Rank-invariant measurement therefore requires rank-invariant γ/r. AR1-ZO uses the rank-1 measurement as the reference: α is the output coefficient assigned to the queried atom, giving $\gamma = \alpha r$. This calibration is topology-specific rather than a revision of joint first-order LoRA scaling; sequential ZO asks one atom, not the joint adapter, to carry the observable directional signal in a two-point measurement.

3.3 Two-Forward-Pass AR1-ZO Update

Algorithm 1 exposes the finest complete LoRA-native atom block and measures it with the topology-aware coefficient. The cyclic schedule visits each atom once per cycle; each step samples the active atom's two factor directions, evaluates paired endpoints, and updates only the selected vectors.

Algorithm 1 AR1-ZO update for one LoRA matrix

Require: Loss $\mathcal{L}$, rank r, active LoRA scale α, smoothing radius μ, step size η_t, atom schedule $k(t) = 1 + (t \bmod r)$.

1: Set $\gamma = \alpha r$ and choose the active atom $k(t)$.
2: Sample $\boldsymbol{z}_b \sim \mathcal{N}(0, \boldsymbol{I}_{d_{\text{out}}})$ and $\boldsymbol{z}_a \sim \mathcal{N}(0, \boldsymbol{I}_{d_{\text{in}}})$; let $\boldsymbol{z}_t = (\boldsymbol{z}_b, \boldsymbol{z}_a)$.
3: Evaluate $\ell_+ = \mathcal{L}(\theta_t + \mu \boldsymbol{P}_t \boldsymbol{z}_t)$ and $\ell_- = \mathcal{L}(\theta_t - \mu \boldsymbol{P}_t \boldsymbol{z}_t)$ on a paired mini-batch.
4: Form the active-atom estimator $\hat{\boldsymbol{g}}_t = ((\ell_+ - \ell_-)/(2\mu))\, \boldsymbol{z}_t$ and split it as $(\hat{\boldsymbol{g}}_{b,t}, \hat{\boldsymbol{g}}_{a,t})$.
5: Update $\boldsymbol{b}_{k(t)} \leftarrow \boldsymbol{b}_{k(t)} - \eta_t \hat{\boldsymbol{g}}_{b,t}$ and $\boldsymbol{a}_{k(t)} \leftarrow \boldsymbol{a}_{k(t)} - \eta_t \hat{\boldsymbol{g}}_{a,t}$; keep all other atoms frozen.

In transformers, the same active rank index is shared across adapted matrices, so one pair of forward passes suffices. A full cycle covers all atoms; the price of rank is delayed coverage, not a larger per-step query budget.

4 Theoretical Analysis

AR1-ZO changes the role of LoRA rank in a two-point finite-difference measurement. A rank-r adapter still stores r atoms, but a single query perturbs only one matched atom and observes it through the LoRA coefficient γ/r in Eq. (1). Three quantities govern the resulting estimator: the dimension of the active random direction, the signal scale of the active finite difference, and the time needed to revisit all stored atoms. Matched atoms give a complete LoRA-native query block of dimension $q = d_{\text{out}} + d_{\text{in}}$ rather than rq; inherited joint scaling makes the active numerator vanish with rank; topology-aware scaling restores rank-invariant active signal, leaving rank only as an amortized coverage cost. When an active atom aligns with dense gradient structure, the same low-dimensional query can also carry dense descent signal. The rank-invariance claim refers to the active finite-difference measurement, conditional on the active-atom gradient regularity formalized in the appendix; it is logically separate from the forward-scale stability used by the convergence bound. Complete proofs appear in Appendix C.

4.1 Rank-1 Atoms as Minimal LoRA-Native Query Blocks

The matched atom is the smallest complete block in the LoRA factors. Perturbing only $\boldsymbol{b}_k$ or only $\boldsymbol{a}_k$ misses first-order variations of $\boldsymbol{b}_k \boldsymbol{a}_k^\top$, while perturbing $(\boldsymbol{b}_k, \boldsymbol{a}_k)$ keeps all local factor directions of that atom inside the original LoRA parameterization. AR1-ZO is therefore a LoRA-native atom query, not an auxiliary subspace or one-sided factor shortcut.

Lemma 4.1 (Completeness and minimality of matched rank-1 atoms)**.** *For nonzero factors, one-sided blocks are incomplete. The matched atom $(\boldsymbol{b}_k, \boldsymbol{a}_k)$ is minimal complete, has dimension $q = d_{\text{out}} + d_{\text{in}}$, and extends to mq for rank-m blocks.*

Minimality fixes the smallest complete LoRA-native coordinates available to the oracle. The direction $\boldsymbol{z}$ has only q atom coordinates and is lifted by the active projector; all other atoms are identical at both endpoints. Under L-smoothness, projected RGE moment bounds, and residual variance σ_ξ^2, the two-point error bound counts the sampled atom dimension, not the stored rank-r adapter.

Proposition 4.2 (Active-atom ZO dimension is independent of rank)**.** *For $k = k(t)$,*

$$\mathbb{E}\Big[\|\hat{\nabla}_k\mathcal{L} - \nabla_k\mathcal{L}(\theta)\|^2\Big] \lesssim q\|\nabla_k\mathcal{L}(\theta)\|^2 + L^2\mu^2q^3 + \frac{\sigma_\xi^2 q}{\mu^2}.$$

No term contains rq.

These statements separate stored capacity from instantaneous sampling dimension. Rank controls atom coverage, not the variance dimension of one query. The remaining issue is scale: endpoint losses see the active atom only after multiplication by γ/r.

4.2 Diagnosis: Directional Collapse under Naive Scaling

Sequential rank-1 querying exposes the active gradient through the coefficient γ/r. Fix iteration t and write $k = k(t)$. Let $\boldsymbol{g}_t = [\boldsymbol{G}\boldsymbol{a}_k; \boldsymbol{G}^\top\boldsymbol{b}_k]$ be the unscaled active factor-coordinate gradient, where $\boldsymbol{G} = \nabla_{\boldsymbol{W}}\mathcal{L}$ is the dense matrix gradient; the gradient actually seen by the two-point oracle is $(\gamma/r)\boldsymbol{g}_t$. Since the oracle receives only a noisy scalar projection of this vector, the active FD-SNR, denoted SNR_t below, determines whether the direction remains observable. The numerator of Eq. (4) gives

$$\frac{\Delta_\mu(\theta; \boldsymbol{P}_t\boldsymbol{z}) + \xi_+ - \xi_-}{2\mu} = \frac{\gamma}{r}\langle\boldsymbol{g}_t, \boldsymbol{z}\rangle + \frac{\xi_+ - \xi_-}{2\mu} + \mathcal{O}(\mu^2), \quad \mathsf{SNR}_t \asymp \frac{(\gamma/r)^2\mu^2\|\boldsymbol{g}_t\|^2}{\sigma_\xi^2}. \tag{7}$$

This active FD-SNR belongs to the measured function difference, not the subsequent optimizer step. With joint LoRA scaling $\gamma = \alpha$, the clean numerator is divided by r while residual endpoint noise is not. High rank can therefore erase directional content before any learning rate is applied.

Theorem 4.3 (Directional Collapse under Naive Scaling)**.** *Naive $\gamma = \alpha$ gives active signal $\Theta(r^{-1})$ and active FD-SNR $\mathsf{SNR}_t = \Theta(r^{-2})$. With $r_c = \Theta(\alpha\mu\|\boldsymbol{g}_t\|/\sigma_\xi)$,*

$$\frac{r}{r_c} \to \infty \quad \Longrightarrow \quad \mathbb{E}\Big[\cos\Big(\hat{\nabla}_k\mathcal{L}, \nabla_k\mathcal{L}\Big)\Big] \to 0.$$

Directional collapse is a measurement failure, not a lack of adapter capacity. A low-dimensional active atom can still be drowned by residual noise once α/r suppresses the clean numerator. A larger learning rate only rescales the already noisy scalar difference, so useful high-rank updates require a query-time correction of the LoRA coefficient.

4.3 Repair: Signal Restoration and Coverage Cost

Topology-aware scaling applies the correction before endpoint losses are queried. Setting $\gamma = \alpha r$ makes $\gamma/r = \alpha$, so the clean numerator in Eq. (7) no longer decays with rank. The FD-SNR claim uses full $\Theta(1)$ active-atom gradient regularity; the convergence claim uses stable represented-adapter scale to rule out output inflation and the upper-bound half $\|\boldsymbol{g}_t\| = \mathcal{O}(1)$ to prevent hidden rank dependence in the descent inequality. The appendix states these complementary assumptions separately.

Theorem 4.4 (Signal Restoration and Active-Atom Stationarity)**.** *With $\gamma = \alpha r$: **(i)** active FD-SNR is rank-invariant, $\mathsf{SNR}_t \asymp \alpha^2\mu^2\|\boldsymbol{g}_t\|^2/\sigma_\xi^2 = \Theta(1)$; **(ii)** the projected RGE convergence bound holds,*

$$T^{-1}\sum\nolimits_{t=1}^{T}\mathbb{E}\|\nabla_{k(t)}\mathcal{L}(\theta_t)\|^2 \le \mathcal{O}(T^{-1/2}) + \mathcal{O}\left(\mu^4L^2q^2 + \frac{\sigma_\xi^2 q}{T^{1/2}\mu^2}\right).$$

Part (i) uses active-atom gradient regularity; part (ii) uses forward-scale stability of the represented adapter.

The bound depends on q, residual noise σ_ξ, and smoothing scale, but not on rq or the naive r^{-2} active FD-SNR decay. Rank reappears only through atom coverage:

Corollary 4.5 (Coverage cost for full-adapter stationarity). *For uniform atom sampling, or a cyclic schedule averaged over one cycle,*

$$\mathbb{E}_{k(t)}\left[\|\nabla_{k(t)}\mathcal{L}(\theta)\|^2\right] = \frac{1}{r}\sum_{k=1}^{r}\|\nabla_k\mathcal{L}(\theta)\|^2.$$

Full-adapter stationarity therefore costs atom coverage, not a larger per-query dimension.

The same calibration applies to wider blocks: a rank-m active query trades higher per-step dimension mq for a shorter coverage cycle.

Corollary 4.6 (Extension to Rank-m Blocks and Other Factorizations). *A rank-m active block with target coefficient α uses $\gamma = \alpha r/m$ and query dimension mq. The same compensation applies to joint-normalized multilinear factorizations under sequential block queries.*

4.4 Structural Insight: Why Native Atoms Can Carry Dense Descent

A low-dimensional measurement is useful only when the measured atom carries dense descent signal. Let $\boldsymbol{G} = \nabla_{\boldsymbol{W}}\mathcal{L} = \sum_i \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, define $\beta = \cos^2(\boldsymbol{b}_k, \boldsymbol{u}_1)\cos^2(\boldsymbol{a}_k, \boldsymbol{v}_1)$ as alignment with the top singular pair, and write $\boldsymbol{g}_k = [\boldsymbol{G}\boldsymbol{a}_k; \boldsymbol{G}^\top\boldsymbol{b}_k]$ for the unscaled active gradient. For $\Delta\mathcal{L} = \mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t)$, the one-step bound first depends on active gradient energy and then relates that energy to spectral concentration and atom alignment.

Theorem 4.7 (Conditional Progress Advantage of Aligned Rank-1 Atoms). *If $\sigma_1^2 \geq \rho\|\boldsymbol{G}\|_F^2$, then*

$$\mathbb{E}[-\Delta\mathcal{L}] \gtrsim \eta\left(\frac{\gamma}{r}\right)^2\|\boldsymbol{g}_k\|^2 - \frac{\eta^2 L}{2}\mathbb{E}\left[\|\hat{\nabla}_k\mathcal{L}\|^2\right] - \mathcal{O}(\eta\mu^4 L^2 q^2),$$
$$\|\boldsymbol{g}_k\|^2 \geq \rho\beta\|\boldsymbol{G}\|_F^2(\|\boldsymbol{a}_k\|^2 + \|\boldsymbol{b}_k\|^2).$$

Alignment need not appear in every layer or at every training stage. When it is present, a single intrinsic rank-1 query carries a $\rho\beta$-scaled portion of dense descent signal while retaining active dimension q. The active-atom factor-norm regularity used in Section 4.3 turns $(\|\boldsymbol{a}_{k(t)}\|^2 + \|\boldsymbol{b}_{k(t)}\|^2)$ into a $\Theta(1)$ multiplier; the signal advantage at active dimension q is then $\Theta(\rho\beta\, d)$ when $d_{\text{out}} = d_{\text{in}} = d$, with the exact ratio against generic random-block estimators given in Appendix C. Together with Theorem 4.4, this separates the two ingredients behind faster practical convergence: q controls the estimator variance, and $\rho\beta$ controls the useful signal. The experiments therefore measure spectral concentration and atom alignment, test the predicted active FD-SNR restoration under rank sweeps, and then evaluate whether restored high-rank capacity improves downstream performance under the same two-forward-pass budget.

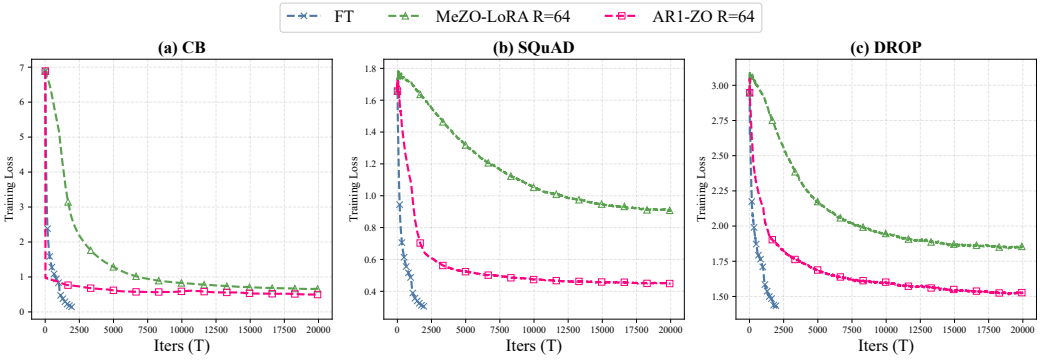


Figure 2: Training-loss trajectories on CB, SQuAD, and DROP. Under matched two-forward budgets, AR1-ZO descends smoothly among ZO methods across both classification and generation-style tasks; FT is shown as a first-order reference.

5 Experiments

The empirical study follows the theory's causal chain: Section 5.2 verifies end-to-end stability on representative CB/SQuAD/DROP loss curves; Section 5.3 opens the box on the spectral premises of

Theorem 4.7 and the rank-signal restoration of Theorems 4.3–4.4; Section 5.4 measures downstream payoff on OPT and Qwen3 families under matched two-forward budgets, with SQuAD controls in Appendix B.2.

5.1 Experimental Setup

Models, tasks, and baselines. Downstream evaluation uses OPT-2.7B/13B as the primary family and Qwen3-1.7B/32B as modern-model confirmation on BoolQ, CB, COPA, WIC, SQuAD, and DROP. OPT experiments compare AR1-ZO against MeZO-LoRA, LOZO, ZO-Alt-naive, and FT (Adam); Qwen3 experiments isolate the scaling axis with MeZO, ZO-Alt-naive, and AR1-ZO. HI-ZOO public-implementation runs are reported in Appendix Table 2 as a reproducibility check under the same evaluation contract. AR1-ZO uses $\alpha = 16$, $\gamma = \alpha r$, and all ZO methods use two forward evaluations per step.

Diagnostic and ablation runs. Mechanism/rank-signal diagnostics run on Qwen3-1.7B/COPA and SQuAD controls move to Appendix B.2. Detailed splits, baseline tuning, hyperparameters, and profiling protocols are in Appendix B.

5.2 Training Dynamics and Convergence Speed

With matched two-forward budgets fixed, we first ask whether AR1-ZO's atom accounting yields *stable end-to-end optimization*, before opening the box in Section 5.3. Theorem 4.4 predicts stable active directions at dimension q rather than rq; if so, loss curves should descend smoothly among matched-budget ZO methods on representative CB, SQuAD, and DROP runs, with FT (Adam) serving as a first-order reference rather than a rate target.

Figure 2 shows that dimension reduction alone does not stabilize the finite-difference measurement across these tasks. AR1-ZO instead cycles atoms and reads each through the rank-invariant coefficient γ/r (the projected-RGE second moment of Theorem 4.4), giving the fastest matched-budget ZO curves and narrowing, but not closing, the FT gap. The full eight-task loss comparison appears in Appendix Fig. 5.

5.3 Mechanism and Rank-Signal Validation

Stable loss curves are necessary but not sufficient, they could arise from any well-tuned ZO scheme. To attribute the gain to our mechanism, we test the two claims behind Theorems 4.7–4.4 in turn: (i) the structural premises (spectral concentration, learned-atom alignment) hold along training, and (ii) the rank-scaling correction $\gamma = \alpha r$ restores active FD-SNR that naive scaling destroys.

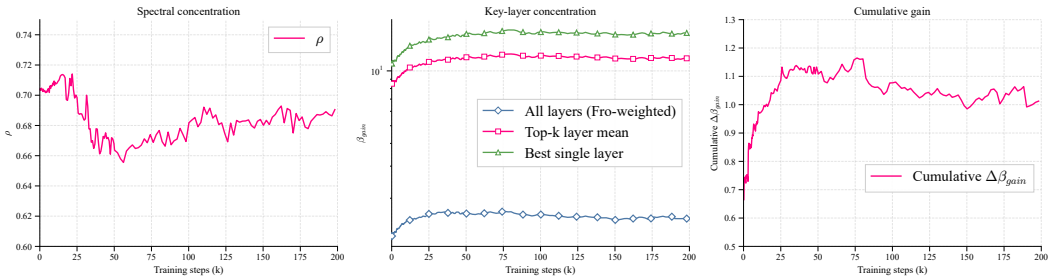


Figure 3: Mechanism validation over 2×10^5 steps. **Left:** Spectral concentration ρ remains robust. **Middle:** Alignment gain concentrates heavily in high-impact layers. **Right:** Cumulative alignment gain remains strictly positive, confirming a net structural pull.

Claim (i): the gradient lives in a small structured subspace, and AR1-ZO aligns with it. On Qwen3-1.7B/COPA over 2×10^5 steps, the three panels of Fig. 3 give complementary readings. Spectral concentration $\rho \in [0.65, 0.75]$ *(left)* means a few directions carry most of the gradient; top-2 layer alignment at 10–30$\times$ the random null *(middle)* means AR1-ZO's atoms hit those directions in the high-impact layers, not random ones; the overall cumulative alignment gain in $[2, 7]$ *(right)*, against zero for memoryless random projections, means the alignment compounds rather than cancels across steps. Appendix B.1 reports the $\boldsymbol{b}$- and $\boldsymbol{a}$-side cosine decomposition.

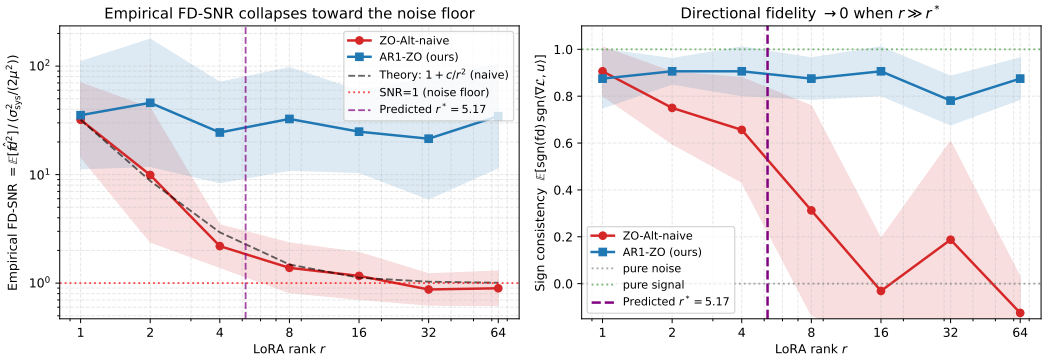


Figure 4: Signal restoration validation. Active FD-SNR and directional fidelity collapse under naive scaling but remain stable under AR1-ZO.

Claim (ii): naive scaling starves the signal; $\gamma = \alpha r$ feeds it back. Sweeping $r \in \{1, 4, 8, 16, 32, 64\}$ on the same alternating topology, Fig. 4 reads two health metrics: active FD-SNR (the signal-to-noise the optimizer can actually use) and directional fidelity (cosine to the true gradient). Naive scaling (ZO-Alt-naive) collapses both—FD-SNR drops below the optimization threshold at $r = 64$ and directional fidelity decays to pure noise—exactly the topology-scaling mismatch of Theorem 4.3. With $\gamma = \alpha r$, AR1-ZO holds both metrics steady across the entire sweep, so capacity can grow with r without destroying what AR1-ZO measures (Theorem 4.4).

Table 1: **Main downstream results.** Accuracy (%) for classification and F1 for SQuAD/DROP; differences are relative to MeZO-LoRA.

OPT-2.7B

Method	BoolQ	CB	Copa	WIC	SQuAD	DROP
MeZO-LoRA (r=64)	62.2	66.1	72.0	59.7	64.6	23.9
LOZO	**65.6** (+3.4)	67.9 (+1.8)	**80.0** (+8.0)	54.4 (-5.3)	76.2 (+11.6)	23.5 (-0.4)
ZO-Alt-Naive	55.5 (-6.7)	28.6 (-37.5)	71.0 (-1.0)	54.9 (-4.8)	19.2 (-45.4)	6.3 (-17.6)
AR1-ZO (r=64)	64.6 (+2.4)	**69.6** (+3.5)	79.0 (+7.0)	**60.3** (+0.6)	**77.8** (+13.2)	**27.8** (+3.9)
FT (Adam)	73.3 (+11.1)	85.7 (+19.6)	81.0 (+9.0)	63.8 (+4.1)	84.2 (+19.6)	31.4 (+7.5)

OPT-13B

Method	BoolQ	CB	Copa	WIC	SQuAD	DROP
MeZO-LoRA (r=64)	64.8	66.1	84.0	54.9	78.7	30.1
LOZO	68.1 (+3.3)	**69.6** (+3.5)	**86.0** (+2.0)	56.9 (+2.0)	**81.4** (+2.7)	30.6 (+0.5)
ZO-Alt-Naive	56.9 (-7.9)	33.7 (-32.4)	75.0 (-9.0)	55.8 (+0.9)	20.7 (-58.0)	8.1 (-22.0)
AR1-ZO (r=64)	**66.7** (+1.9)	**69.6** (+3.5)	**86.0** (+2.0)	**57.5** (+2.6)	79.4 (+0.7)	**30.8** (+0.7)
FT (Adam)	76.9 (+12.1)	84.1 (+18.0)	79.0 (-5.0)	70.1 (+15.2)	84.9 (+6.2)	31.3 (+1.2)

Qwen3-1.7B

Method	BoolQ	CB	Copa	WIC	SQuAD	DROP
MeZO (r=64)	74.7	73.2	71.0	53.3	62.0	40.1
ZO-Alt-Naive	68.5 (-6.2)	41.1 (-32.1)	71.0 (+0.0)	49.8 (-3.5)	17.3 (-44.7)	10.2 (-29.9)
AR1-ZO (r=64)	**78.0** (+3.3)	**87.5** (+14.3)	**72.0** (+1.0)	**56.6** (+3.3)	**76.7** (+14.7)	**42.9** (+2.8)

Qwen3-32B

Method	BoolQ	CB	Copa	WIC	SQuAD	DROP
MeZO (r=64)	83.6	89.3	77.0	68.3	73.2	**58.5**
ZO-Alt-Naive	71.6 (-12.0)	45.4 (-43.9)	73.0 (-4.0)	63.3 (-5.0)	21.1 (-52.1)	15.1 (-43.4)
AR1-ZO (r=64)	**85.6** (+2.0)	**90.4** (+1.1)	**88.0** (+11.0)	**71.5** (+3.2)	**80.6** (+7.4)	52.2 (-6.3)

5.4 Capacity Scaling and Downstream Performance

Premises and restored signal in Section 5.3 are means, not ends; the payoff test is whether they translate into matched-budget downstream gains across model families and scales. We fix $r = 64$ for LoRA-based methods and report a two-family two-scale matrix: OPT-2.7B/13B with full ZO baselines plus FT (Adam) as a first-order reference, and Qwen3-1.7B/32B with MeZO, ZO-Alt-naive, and AR1-ZO to isolate the scaling-law axis on a recent model family.

OPT family. Table 1 shows AR1-ZO improves over MeZO-LoRA on all six OPT-2.7B tasks ($+0.6$–$+13.2$) and remains closest to FT (Adam) on COPA (79.0 vs 81.0) and DROP (27.8 vs 31.4); the same pattern of avoiding ZO-Alt-naive collapse holds at OPT-13B on CB/SQuAD/DROP per Fig. 4. **Qwen3 family.** AR1-ZO improves over MeZO by $+1.0$ to $+14.7$ on Qwen3-1.7B, and Qwen3-32B supplies the second modern-model scale. **Controls.** Appendix B.2 shows that sublinear scaling

laws stagnate, learning-rate amplification cannot replace pre-oracle scaling, and AR1-ZO adds no systems overhead.

6 Conclusion

We identified and resolved a topology-scaling mismatch in zeroth-order LoRA fine-tuning: inheriting α/r under sequential atomwise queries drives the active signal below the finite-difference noise floor. **AR1-ZO** pairs alternating rank-1 probes with $\gamma = \alpha r$, making high-rank LoRA usable under pure black-box constraints. Rank leaves the single-query dimension and active FD-SNR, remaining only as amortized atom coverage.

Limitations and Future Work. Theorem 4.4 is active-atom stationarity, so full-adapter stationarity still pays the coverage factor in Corollary 4.5. Appendix B.7–B.8 extends evaluation to DoRA, dialogue summarization, and unseen-task instruction following. MoE models, tensorized adapters, and uniform guarantees beyond the tested rank range remain open.

Acknowledgments and Disclosure of Funding

This work was supported in part by the Youth S&T Talent Support Programme of Guangdong Provincial Association for Science and Technology (Grant No. SKXRC2025466) and Beijing Xiaomi Mobile Software Co., Ltd. (Xiaomi). The funding from Xiaomi was provided through a research collaboration between Xiaomi and The Hong Kong University of Science and Technology (Guangzhou).

References

[1] H. Cai, Y. Lou, D. McKenzie, and W. Yin. A zeroth-order block coordinate descent algorithm for huge-scale black-box optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1193–1203. PMLR, 2021.

[2] S. Chen, Y. Ju, H. Dalal, Z. Zhu, and A. J. Khisti. Robust federated finetuning of foundation models via alternating minimization of loRA. In *Workshop on Efficient Systems for Foundation Models II @ ICML2024*, 2024.

[3] Y. Chen, Y. Zhang, L. Cao, K. Yuan, and Z. Wen. Enhancing zeroth-order fine-tuning for language models with low-rank structures. In *The Thirteenth International Conference on Learning Representations*, 2025.

[4] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc., 2023.

[5] Y. Feng and Y. Shu. ZoAQ: Adaptive zeroth-order querying via query-reuse coupling. *arXiv preprint arXiv:2609.22115*, 2026.

[6] B. Gliwa, I. Mochol, M. Biesek, and A. Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In L. Wang, J. C. K. Cheung, G. Carenini, and F. Liu, editors, *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[7] S. Hayou, N. Ghosh, and B. Yu. LoRA+: Efficient low rank adaptation of large models. In *Forty-first International Conference on Machine Learning*, 2024.

[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[9] D. Kalajdzievski. A rank stabilization scaling factor for fine-tuning with lora, 2023.

[10] W. Lin, Y. Jiang, Q. Song, Q. Xiang, and H. Xu. Agzo: Activation-guided zeroth-order optimization for llm fine-tuning, 2026.

[11] J. Liu, Z. Kong, P. Dong, C. Yang, X. Shen, P. Zhao, H. Tang, G. Yuan, W. Niu, W. Zhang, X. Lin, D. Huang, and Y. Wang. RoRA: Efficient fine-tuning of LLM with reliability optimization for rank adaptation. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[12] S. Liu, B. Kailkhura, P.-Y. Chen, P. Ting, S. Chang, and L. Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[13] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, and M.-H. Chen. Dora: weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[14] S. Malladi, T. Gao, E. Nichani, A. Damian, J. D. Lee, D. Chen, and S. Arora. Fine-tuning language models with just forward passes. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 53038–53075. Curran Associates, Inc., 2023.

[15] Y. Nesterov and V. Spokoiny. Random gradient-free minimization of convex functions. *Found. Comput. Math.*, 17(2):527566, Apr. 2017.

[16] S. Park, J. Yun, S. Kim, S. Kundu, and E. Yang. Elucidating subspace perturbation in zeroth-order optimization: Theory and practice at scale, 2025.

[17] J. Qiu, Z. Hong, R. Xu, and Y. Shu. Revisiting zeroth-order hessian approximation: A single-step policy optimization lens. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 102036–102072. PMLR, 2026.

[18] J. Qiu, Z. Xie, X. Yan, Y. Yang, and Y. Shu. Zeroth-order optimization is secretly single-step policy optimization. In *ICML Workshop on Tiny Titans: The Next Wave of On-Device Learning for Foundational Models*, 2025.

[19] H. Seung, J. Lee, and H. Ko. Low-rank curvature for zeroth-order optimization in LLM fine-tuning, 2025. Accepted to AAAI 2026.

[20] Z. Song and W. Li. RoZO: Geometry-aware zeroth-order fine-tuning on low-rank adapters for black-box large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1773–1783, Rabat, Morocco, 2026. Association for Computational Linguistics.

[21] Y. Sun, T. Huang, L. Ding, L. Shen, and D. Tao. TeZO: Empowering the low-rankness on the temporal dimension in the zeroth-order optimization for fine-tuning LLMs, 2025.

[22] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Naik, A. Ashok, A. S. Dhanasekaran, A. Arunkumar, D. Stap, E. Pathak, G. Karamanolakis, H. Lai, I. Purohit, I. Mondal, J. Anderson, K. Kuznia, K. Doshi, K. K. Pal, M. Patel, M. Moradshahi, M. Parmar, M. Purohit, N. Varshney, P. R. Kaza, P. Verma, R. S. Puri, R. Karia, S. Doshi, S. K. Sampat, S. Mishra, S. Reddy A, S. Patro, T. Dixit, and X. Shen. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.

[23] C. Wei, Y. Shu, Y. T. He, and F. Yu. Flexora: Flexible low-rank adaptation for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14643–14682. Association for Computational Linguistics, 2025.

[24] C. Wei, H. Wang, Y. He, F. Yu, and Y. Shu. Test-time policy adaptation for enhanced multi-turn interactions with LLMs. In *NeurIPS Workshop on Multi-Turn Interactions in Large Language Models*, 2025.

[25] Z. Yu, P. Zhou, S. Wang, J. Li, M. Tian, and H. Huang. Zeroth-order fine-tuning of LLMs in random subspaces, 2025. ICCV 2025 camera-ready version.

[26] Q. Zhang, M. Chen, A. Bukharin, P. He, Y. Cheng, W. Chen, and T. Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*, 2023.

[27] Y. Zhao, S. Dang, H. Ye, G. Dai, Y. Qian, and I. Tsang. Second-order fine-tuning without pain for LLMs: A hessian informed zeroth-order optimizer. In *The Thirteenth International Conference on Learning Representations*, 2025.

A Related Work

Optimizing LLMs under strict memory constraints has driven the convergence of LoRA-style adaptation and zeroth-order (ZO) optimization. The relevant literature is therefore not a single lineage but the intersection of three threads: LoRA's rank-factor structure, dimensionality reduction in ZO, and LoRA scaling dynamics. AR1-ZO is defined by two specific claims rather than by the broad fact that LoRA is low-rank. First, it uses each matched column-row rank-1 atom as the complete query block exposed to a two-point ZO oracle. Second, it shows that this sequential query topology changes the meaning of LoRA's scale and requires the topology-aware correction $\gamma = \alpha r$ to keep the finite-difference signal observable. Query-efficient ZO estimators provide additional context for these two claims.

LoRA and Algebraic Decomposition. LoRA freezes pre-trained weights and injects trainable low-rank modules. Its update $\Delta \boldsymbol{W} = \boldsymbol{B}\boldsymbol{A}$ can be written as an additive ensemble of rank-1 atoms, and structural refinements such as AdaLoRA [26] make related rank components explicit. Because these components are additively separable, they invite alternating and blockwise optimization: RoLoRA [2] alternates $\boldsymbol{A}/\boldsymbol{B}$ to stabilize federated FO training, and MeZO-BCD [16] updates parameter subsets to reduce effective ZO dimensionality. These works establish that LoRA factors and parameter blocks can be exposed selectively, but they do not use the matched rank-1 atom $(\boldsymbol{b}_k, \boldsymbol{a}_k)$ as the minimal complete two-point ZO query block. The distinction matters because a one-sided factor update does not span all local variations of an atom, while larger blocks reintroduce rank-dependent query dimension. At the layer level, Flexora [23] formulates adapter placement as a hyperparameter optimization problem and uses unrolled differentiation to select layers for LoRA fine-tuning. This layer-selection perspective concerns where adaptation capacity is allocated, while atomwise querying concerns the measurement granularity within each selected adapter. ROSA [24] uses low-rank updates for test-time policy adaptation from interaction feedback, including LoRA on the language-model head. Its linearized solver requires Jacobian-vector and vector-Jacobian products, placing it in derivative-based adaptation. *Our positioning:* the atom contribution is not the identity $\boldsymbol{B}\boldsymbol{A} = \sum_k \boldsymbol{b}_k \boldsymbol{a}_k^\top$ itself; it is the use of the complete rank-1 atom as the LoRA-native measurement unit that reduces the per-query dimension without reducing stored adapter rank.

Subspace and Low-Rank Zeroth-Order Optimization. Standard ZO methods such as MeZO [14] suffer variance proportional to the full parameter dimension, prompting several subspace remedies. SubZero [25] restricts perturbations to random low-dimensional subspaces, while LOZO [3] builds low-rank ZO estimators from sampled low-rank perturbations. TeZO [21] extends this low-rank viewpoint across training time by constructing perturbations through a temporal tensor decomposition, and LOREN [19] and HIZOO [27] improve search directions with curvature or Hessian-informed statistics. ZoVH [17] develops this curvature-estimation direction through a single-step policy-optimization interpretation, using an optimal baseline and historical query reuse to construct variance-reduced estimators of the Hessian and related inverse-curvature quantities. AGZO [10] constructs activation-guided perturbation subspaces from forward-pass information. RoZO [20] is the closest LoRA-aware geometry method: it treats LoRA adapters as a low-rank manifold and constrains ZO updates to tangent spaces with Riemannian tools. *Our positioning:* AR1-ZO is complementary to these subspace, low-rank, and geometry methods. It does not

sample an auxiliary basis, use activation hooks, estimate curvature, decompose perturbations over time, or optimize on the full LoRA manifold tangent space. Instead, it queries the structural rank-1 atoms already present in the parameterization. TeZO shows that temporal structure can reduce ZO perturbation overhead, and RoZO shows that LoRA geometry is useful for ZO; AR1-ZO studies the finer measurement topology obtained when one atom is queried at a time, and identifies the scaling pathology that appears only under this sequential atomwise oracle.

Variance Reduction and Query Allocation. The statistical use of function evaluations offers another route to efficient ZO optimization. ZoAR [18] establishes an equivalence between Gaussian-smoothed ZO gradient estimation and single-step policy optimization with a baseline, then uses an averaged baseline and historical query reuse to reduce estimation variance. ZoAQ [5] couples query reuse with adaptive query allocation: previous evaluations support both the update estimate and the decision to acquire further queries, avoiding separate validation queries. These methods study how evaluations are reused and how many are needed for an update. AR1-ZO studies the LoRA-native block exposed to each paired evaluation and the scale needed to preserve its active finite-difference signal under a fixed two-forward-pass budget. This separates improvements in estimator statistics and query allocation from the atomwise measurement calibration developed here.

Optimization Topology and Scaling Dynamics. LoRA scaling laws such as the canonical α/r [8], rsLoRA's $\alpha/\sqrt{r}$ [9], and RoRA's related rank-reliability correction [11] study how adapter magnitude should vary with rank during ordinary LoRA fine-tuning. LoRA+ [7] addresses a different but related optimization imbalance by assigning different learning rates to $\boldsymbol{A}$ and $\boldsymbol{B}$. These works are important because they show that LoRA scaling is not a harmless implementation detail. Their topology, however, is still joint first-order adaptation: all rank components participate in the update signal together. *Our positioning:* AR1-ZO's scaling claim is not merely that a larger scalar can improve high-rank LoRA. The claim is that when a two-point ZO oracle perturbs only one rank-1 atom, the inherited joint coefficient α/r multiplies the clean finite-difference numerator while the residual evaluation noise is unchanged. This makes the active finite-difference signal-to-noise ratio (FD-SNR) decay as $1/r^2$. The correction $\gamma = \alpha r$ is therefore a topology-aware measurement calibration: it assigns the active atom a rank-invariant observable signal while keeping the remaining atoms fixed across the paired evaluations.

B Experimental Details

Comparison roles and contract. The empirical study separates four evidence layers, each carried by its own subsection in the main text. (i) Trajectory-level convergence and stability are measured directly on the primary downstream family (OPT-2.7B). (ii) Mechanism diagnostics on COPA test the theorem-level quantities used by the analysis: spectral concentration ρ, atom alignment gain β_{gain}, active FD-SNR, and directional fidelity. (iii) Downstream tables test whether the restored high-rank signal remains useful under matched two-forward-pass budgets, on the OPT family (2.7B and 13B) as the primary axis and on Qwen3 (1.7B and 32B) as a modern-model confirmation. (iv) SQuAD ablations isolate scaling law choice, rule out learning-rate compensation, and measure step time and peak memory under the same profiling protocol. Keeping these four axes apart prevents downstream accuracy from being treated as the only evidence and lets each theorem be tested on the quantity it actually predicts.

Tasks, splits, and models. The main downstream evaluation uses OPT-2.7B, OPT-13B, Qwen3-1.7B, and Qwen3-32B on BoolQ, CB, COPA, WIC, SQuAD, and DROP. The full loss-trajectory diagnostic additionally includes WSC and ReCoRD, reported only in Appendix Fig. 5. The explicit split sizes used in the main text are CB $(250/56/250)$, COPA $(400/100/500)$, and SQuAD $(1000/500/1000)$ for the ablation setting. Mechanism and rank-diagnostic experiments use COPA, while scaling-law and efficiency controls use SQuAD with Qwen3-1.7B.

Baselines and query matching. On the OPT family AR1-ZO is compared against unstructured ZO (MeZO, MeZO-LoRA), structured ZO (LOZO), the diagnostic ZO-Alt-naive baseline, and Adam full fine-tuning (FT) as a first-order upper bound from the MeZO comparison setting. HIZOO public-implementation runs are reported separately in Table 2 as a reproducibility check because the released code showed unstable performance under the matched LoRA-ZO evaluation contract. On

Table 2: **HIZOO public-code reproducibility check.** Results are from the released HIZOO implementation under the same OPT LoRA-ZO evaluation contract used in the main table; differences are relative to MeZO-LoRA.

Model	BoolQ	CB	COPA	WIC	SQuAD	DROP
OPT-2.7B	60.8 (-1.4)	55.4 (-10.7)	60.0 (-12.0)	45.6 (-14.1)	0.0 (-64.6)	0.1 (-23.8)
OPT-13B	63.1 (-10.4)	41.0 (-25.1)	52.0 (-32.1)	50.0 (-10.2)	0.0 (-82.1)	0.1 (-30.6)

the Qwen3 family the comparison narrows to MeZO, ZO-Alt-naive, and AR1-ZO, isolating the scaling-law axis from auxiliary-basis or curvature-side-information effects. ZO-Alt-naive uses the same alternating rank-1 topology as AR1-ZO but keeps the canonical joint-update scaling, making it the direct control for the topology-scaling correction. All ZO methods are matched to two forward evaluations per optimization step.

Baseline tuning and fairness. All reported baselines use the same task splits, batch size, evaluation budget, and two-forward-pass ZO step contract as AR1-ZO. LOZO and HIZOO use the best hyperparameter settings from their released GitHub scripts rather than AR1-ZO-specific rank or scaling choices; HIZOO is separated into Table 2 because that public implementation was unstable under the matched LoRA-ZO contract. MeZO-LoRA also follows the best script hyperparameters, with LoRA rank and α set to 64 for the high-rank comparison. ZO-Alt-naive is not tuned as an independent competitor: it shares AR1-ZO's alternating rank-1 query topology and optimizer settings but keeps canonical joint LoRA scaling, making it the direct diagnostic control for topology-scaling mismatch. TeZO and RoZO are discussed as related methods but are not included as empirical baselines because no public implementation was available at the time of evaluation.

Diagnostic protocols. For the mechanism validation in Fig. 3, Qwen3-1.7B is fine-tuned with LoRA rank $r = 64$ on COPA for $2{\times}10^5$ steps. The full 400-example training split is used as a fixed diagnostic probe set, and spectral/alignment statistics are aggregated across the 28 v_{proj} matrices with Frobenius weighting. For the rank sweep in Fig. 4, LoRA rank varies over $r \in \{1, 4, 8, 16, 32, 64\}$, the function-evaluation noise σ_ξ is isolated through controlled variance injection, and the finite-difference probes use unit-normalized Gaussian perturbations.

Ablation and efficiency protocols. The SQuAD ablations in Fig. 7 use Qwen3-1.7B, LoRA rank $r = 64$, $\alpha = 16$ unless otherwise noted, batch size 16, smoothing radius $\mu = 10^{-3}$, and seed 42. The scaling-law sweep compares active output coefficients $\gamma/r \in \{\alpha/r, \alpha/\sqrt{r}, \alpha\}$, equivalently $\gamma \in \{\alpha, \alpha\sqrt{r}, \alpha r\}$, over $2{\times}10^4$ ZO steps with the alternating topology fixed. The learning-rate control places ZO-Alt-naive in its constant-γ worst case with $\alpha = 1$, $r = 64$, and $\gamma = 1$, then sweeps $\eta \in \{10^{-5}, 10^{-4}, 10^{-3}\}$. Runtime and memory are profiled over 500 ZO steps; LOZO uses lazy refresh 50.

Compute resources. The primary OPT-family experiments are run on NVIDIA A800 GPUs, and the primary Qwen3-family experiments are run on NVIDIA RTX PRO 6000 GPUs. The additional time-to-target and component-level memory profiles use Jetson Thor and H100 hardware, respectively, as specified in Appendix B.6. The efficiency panel in Fig. 7 reports step time and peak VRAM under the same profiling protocol over 500 ZO steps, giving the compute and memory envelope needed to reproduce the matched-budget runs.

B.1 Mechanism Validation Details

This section expands Section 5.3 on the mechanism diagnostics summarized in Fig. 3.

Layer anisotropy. The Frobenius-weighted network average of the alignment gain β_{gain} is modest (1–$3\times$ over the analytical random null), but the dispersion across the 28 v_{proj} matrices is heavy-tailed: the top-2 layer mean reaches 10–$30\times$ and the best single layer reaches 20–$100\times$. The structural advantage of intrinsic atoms therefore concentrates in a task-dependent subset of high-impact layers rather than appearing as a uniform layerwise improvement, consistent with the conditional nature of Theorem 4.7.

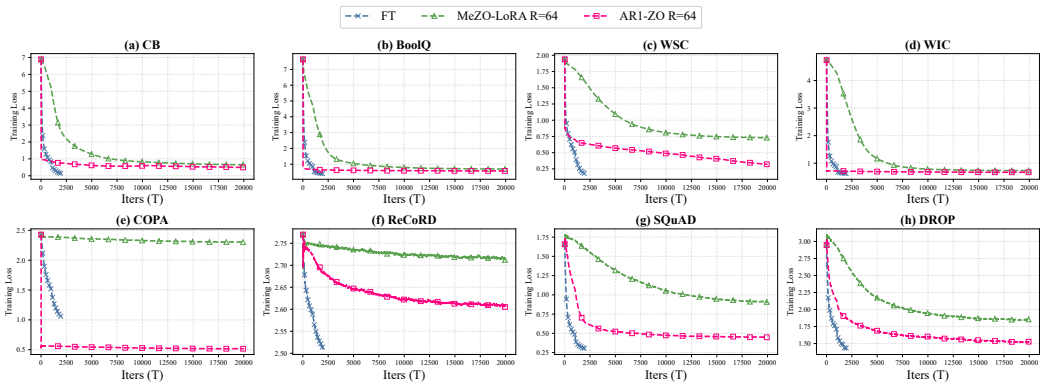


Figure 5: Full eight-task training-loss comparison. The appendix reports the complete loss trajectories on CB, BoolQ, WSC, WIC, COPA, ReCoRD, SQuAD, and DROP; the main text shows a compact three-task subset for readability.

Per-cycle saturation and the cumulative diagnostic. Each atom saturates within its first rank cycle, so per-window β_{gain} increments oscillate around zero in the mid-to-late training regime and do not by themselves expose the directional pull. We therefore track the cumulative quantity $\beta_{\text{gain}}(t) - \beta_{\text{gain}}(t_0)$ against a fixed first-end baseline t_0 chosen at the end of the first complete rank cycle. This trajectory is strictly positive after $\sim 10^4$ steps and stabilizes in $[2, 7]$, while a memoryless random projection (e.g. LOZO's sampled low-rank perturbation) has zero expected cumulative gain under the same diagnostic. The diagnostic therefore distinguishes intrinsic-atom locking from random-block sampling at the trajectory level rather than the per-step level.

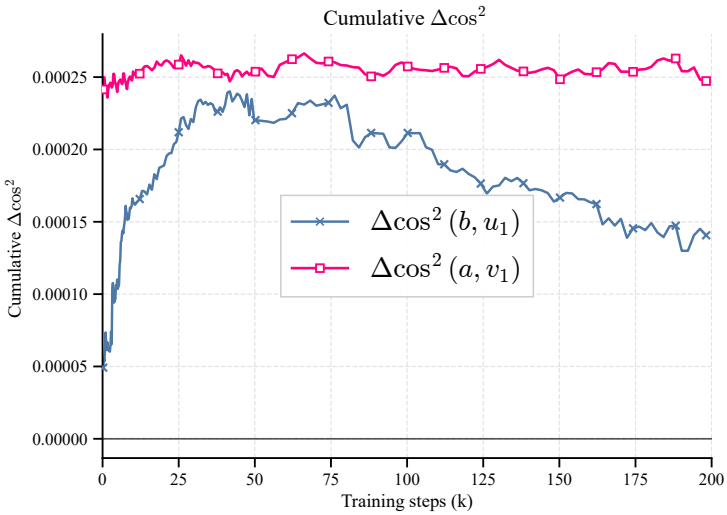


Figure 6: Side-specific cumulative alignment decomposition. The output-side increment $\Delta\cos^2(\boldsymbol{b}, \boldsymbol{u}_1)$ contributes most of the positive cumulative alignment, while the input-side increment $\Delta\cos^2(\boldsymbol{a}, \boldsymbol{v}_1)$ remains smaller but positive.

***b*- and *a*-side decomposition.** Figure 6 decomposes $\beta = \cos^2(\boldsymbol{b}, \boldsymbol{u}_1)\cos^2(\boldsymbol{a}, \boldsymbol{v}_1)$ and shows that the cumulative gain is driven mainly by the output-side cosine $\cos^2(\boldsymbol{b}, \boldsymbol{u}_1)$, with a smaller positive contribution from the input-side $\cos^2(\boldsymbol{a}, \boldsymbol{v}_1)$. This asymmetry is consistent with the observation that LoRA's $\boldsymbol{B}$ factor (initialized at zero and grown by the optimizer) carries the primary alignment dynamics, while $\boldsymbol{A}$ (initialized random) provides a near-isotropic input-side contribution.

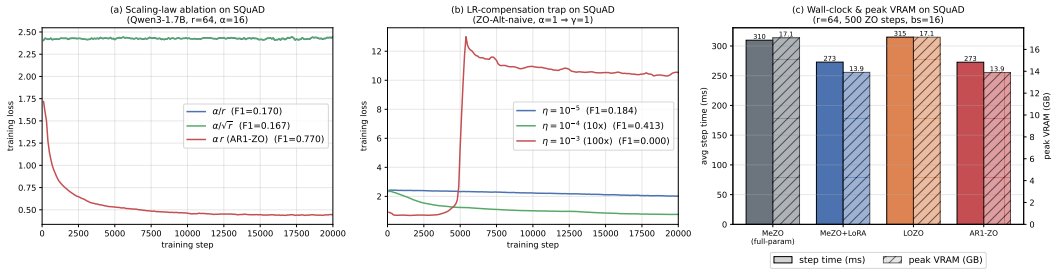


Figure 7: Ablations and computational efficiency on SQuAD. **Left:** Only $\gamma{=}\alpha r$ escapes directional collapse (F1 0.770 vs. ~ 0.17). **Middle:** Amplifying η on ZO-Alt-naive partially recovers ($10\times$, F1 0.413) before destabilizing ($100\times$, F1 0.000). **Right:** AR1-ZO matches MeZO+LoRA exactly (273 ms / 13.9 GB), while LOZO incurs $\sim 15\%$ slower steps and $\sim 23\%$ higher peak VRAM.

B.2 Control Ablation Details

This section expands the controls summarized in Fig. 7.

Scaling-law sweep. Both the canonical active coefficient α/r (the $\gamma = \alpha$ regime of Theorem 4.3) and the milder $\alpha/\sqrt{r}$ stagnate at the noise floor (F1 ≈ 0.17, indistinguishable from random). This confirms that any sub-linear $\gamma(r)$ leaves rank-dependent finite-difference attenuation. In contrast, the rank-invariant active coefficient α implemented by $\gamma = \alpha r$ reaches F1 $= 0.770$ with monotone, non-oscillatory descent, in agreement with the restoration predicted by Theorem 4.4.

Learning-rate compensation trap. The LR-compensation control tests the natural objection that γ is interchangeable with η. It is not: a $10\times$ learning-rate increase on ZO-Alt-naive raises F1 only to 0.413 before training destabilizes around step $5{,}000$, while $100\times$ diverges outright. The asymmetry is structural. The scale γ changes the active atom's output contribution before the finite difference is measured, whereas η amplifies signal and noise after measurement; no scalar learning-rate multiplier can undo the $1/r^2$ active FD-SNR decay diagnosed by Theorem 4.3.

Systems profile. The profiling panel checks that the gains are not purchased with extra systems cost. AR1-ZO matches MeZO+LoRA in step time and peak VRAM (273 ms / 13.9 GB), while LOZO incurs about 15% slower steps and 23% higher peak VRAM. Thus the alternating rank-1 schedule and topology-aware rescaling preserve the pure black-box two-forward-pass budget.

B.3 Query Topology, Block Size, and Rank Coverage

Separating query topology from scale. AR1-ZO changes which factor coordinates are submitted to the two-point oracle: full-adapter MeZO-LoRA queries all r atoms, with dimension rq, whereas AR1-ZO queries one matched atom, with dimension $q = d_{\text{out}} + d_{\text{in}}$. Both retain the same rank-r parameterization and use two scalar loss evaluations. Dimension reduction is consistent with the motivation of subspace and block-coordinate ZO methods [25, 1], but the inherited LoRA coefficient must also be calibrated for the active support. Table 3 separates these two choices on Qwen3-1.7B/SQuAD at $r = 64$ and $2{,}000$ steps. Rank-1 querying with canonical scaling reaches only 17.01 F1, while amplifying all full-adapter coordinates jointly gives 0.00 F1. Their topology-matched combination reaches 58.49 F1. Thus neither component alone reproduces the observed gain. The scale γ acts before endpoint evaluation, changing the measured finite difference; a learning rate acts after measurement and cannot recover directional information lost to residual evaluation noise. The algorithmic contribution is this combination of matched-atom query support, pre-oracle calibration, and cyclic coverage, using the standard two-point estimator.

Complete rank-m blocks and the scope of minimality. Lemma 4.1 concerns complete LoRA-native factor-coordinate blocks. The local variation of atom k is $\delta \boldsymbol{b}_k \boldsymbol{a}_k^\top + \boldsymbol{b}_k \delta \boldsymbol{a}_k^\top$; querying only one factor omits one class of these variations. A complete matched pair therefore contains q coordinates, and grouping m such pairs increases the active dimension to mq. This is a statement about the

Table 3: **Query topology $\times$ scaling.** Qwen3-1.7B/SQuAD, stored rank 64, $2{,}000$ steps, and two forward evaluations per ZO step. F1 is on a 0–100 scale.

Query support	Scale γ	Active coefficient	F1
Full adapter	α	α/r	18.88
Full adapter	αr	α on all atoms	0.00
One matched atom	α	α/r	17.01
One matched atom (AR1-ZO)	αr	α on active atom	**58.49**

native factor blocks, not global optimality over arbitrary partitions, auxiliary bases, activation-guided directions, or data-dependent subspaces. To test the corresponding family, Table 4 varies m while holding the stored rank at 64, the training budget at $2{,}000$ steps, the learning rate at 10^{-5}, and the query budget at two forwards. The cyclic schedule dwells for $h = 200$ steps on each block and uses block-matched scaling $\gamma = \alpha r/m$ with $\alpha = 16$. These runs use batch size 8, smoothing radius 10^{-3}, and seed 42.

Table 4: **Complete rank-m query blocks.** Qwen3-1.7B/SQuAD at stored rank 64. Coverage counts distinct atoms visited by step $2{,}000$; effective rank is averaged across adapted matrices.

Active atoms m	Atoms covered	Mean r_{eff}	SQuAD F1
1	10/64	6.60	**58.49**
2	20/64	15.85	47.35
4	40/64	35.03	30.81
8	64/64	56.44	24.13
16	64/64	59.74	19.69
32	64/64	61.53	18.98
64	64/64	61.53	18.41

For each update matrix $\boldsymbol{BA}$, spectral effective rank is $r_{\text{eff}} = \exp(-\sum_i p_i \log p_i)$, with $p_i = \sigma_i^2 / \sum_j \sigma_j^2$ and the convention $0 \log 0 = 0$. It approaches one for energy concentrated in one singular direction and equals k for equal energy over k directions. Larger blocks cover more atoms and produce higher effective rank within the fixed budget, but F1 decreases as two losses must estimate a direction in more coordinates. The $m = 64$ endpoint reaches 18.41 F1, close to the independently run full-adapter MeZO-LoRA control at 18.88 F1 in Table 3. These results support the single-atom choice within the tested complete-block family; they do not establish universal superiority over other query geometries.

Benefits across stored ranks. Table 5 varies the adapter rank on Qwen3-1.7B/SQuAD under the same $2{,}000$-step, two-forward budget. At $r = 1$, the full-adapter query is already one atom, so AR1-ZO has no dimensional advantage; the observed difference is only 0.94 F1. The gain grows to 12.18, 31.05, and 33.58 points at ranks 4, 8, and 16, and to 39.61 points at rank 64. MeZO-LoRA declines from 65.22 to 18.88 F1 across the sweep, whereas AR1-ZO stays between 54.45 and 66.16 F1. This finite-budget result shows that the benefit is present at intermediate ranks as well as at rank 64. The rank-64 Qwen3-1.7B adapter contains approximately 0.75% as many parameters as the base model.

Table 5: **Stored-rank sweep.** Qwen3-1.7B/SQuAD, $2{,}000$ steps and two forward evaluations per step for both methods. Differences are F1 points.

Stored rank r	MeZO-LoRA	AR1-ZO	Difference
1	65.22	66.16	$+0.94$
2	56.97	63.48	$+6.51$
4	45.37	57.55	$+12.18$
8	27.98	59.03	$+31.05$
16	20.87	54.45	$+33.58$
32	18.55	55.94	$+37.39$
64	18.88	58.49	$+39.61$

Slow coverage and simultaneous updates. All stored atoms contribute to each forward pass; only the active atom changes between the paired evaluations. Proposition 4.2 replaces rq by q in the single-query estimation bound, while Theorem 4.7 gives a conditional $\rho\beta_{\text{gain}}$-fraction of dense descent signal when the active atom is aligned. Neither statement implies that one atom suffices for full-adapter stationarity. Table 6 deliberately delays coverage with $h = 1{,}000$ at ranks 64 and 128. AR1-ZO exceeds full-adapter MeZO-LoRA in all 16 task–rank–checkpoint comparisons, including checkpoints before more than one atom has been trained. At rank 128 and step 500, the gains are 4.1 WIC and 21.9 SQuAD points despite visiting only one atom. This illustrates how a lower-dimensional estimate can be useful before full coverage, without reducing the stored adapter capacity.

Table 6: **Deliberately slow atom coverage.** Qwen3-1.7B with $h = 1{,}000$ steps per atom. Each task cell reports AR1-ZO / MeZO-LoRA; WIC uses accuracy and SQuAD uses F1, both on a 0–100 scale.

Rank	Steps	Trained atoms	WIC	SQuAD
64	500	1/64	54.5 / 49.8	42.8 / 16.8
64	1,000	1/64	54.9 / 49.8	63.4 / 17.1
64	2,000	2/64	54.4 / 49.8	65.4 / 17.5
64	5,000	5/64	56.1 / 49.7	70.4 / 18.4
128	500	1/128	53.9 / 49.8	38.8 / 16.9
128	1,000	1/128	55.6 / 49.8	59.1 / 17.2
128	2,000	2/128	55.3 / 49.8	63.1 / 17.3
128	5,000	5/128	55.5 / 49.7	68.3 / 17.8

For blocks of m atoms with m dividing r, a complete cycle takes rh/m steps. In particular, $m = h = 1$ visits every atom once in r steps and keeps per-atom update counts within one of each other. Rank therefore remains an explicit coverage cost even after it is removed from the single-atom query dimension and the explicit active FD-SNR attenuation. The speed and downstream results here concern the stated finite budgets and schedules.

B.4 Empirical Scope of Active-Gradient Regularity

Assumption C.3 requires that the unscaled active gradient $\boldsymbol{g}_t = [(\boldsymbol{G}\boldsymbol{a}_{k(t)})^\top, (\boldsymbol{G}^\top \boldsymbol{b}_{k(t)})^\top]^\top$ remain of order one in rank along the analyzed iterates. It does not require identical gradient norms or downstream performance at every rank. We measure $\|\boldsymbol{g}_t\|_2$ directly on Qwen3-1.7B/WIC for $r \in \{1, 4, 8, 16, 32, 64\}$ using fixed probe batches and the same cyclic scheduling rule. Exact backward is used only at diagnostic checkpoints 100, 500, $1{,}000$, and $2{,}000$; these gradients never participate in AR1-ZO training.

Table 7: **Active-gradient norm across ranks.** The final two columns summarize variation within each checkpoint. The max/min ratio is descriptive, not a criterion for accepting the assumption.

Step	$r=1$	$r=4$	$r=8$	$r=16$	$r=32$	$r=64$	Max/min	Log–log slope
100	0.695	0.768	0.846	0.631	0.840	0.730	1.34	$+0.011$
500	0.655	0.507	0.602	0.392	0.546	0.414	1.67	-0.092
1,000	0.948	0.595	0.654	0.487	0.472	0.700	2.01	-0.104
2,000	1.121	0.614	0.624	0.555	0.459	1.003	2.44	-0.078

Across these probes the norms lie in $[0.392, 1.121]$, with log–log slopes between -0.104 and 0.011 and a pooled slope of -0.066. A hidden $1/r$ decay would instead give a slope near -1 and approximately 64-fold attenuation from rank 1 to 64. Even the most negative measured slope corresponds to $64^{-0.104} \approx 0.65$, and the pooled slope to approximately 0.76. The variation is also non-monotonic: at step $2{,}000$, the norm is 0.459 at rank 32 and 1.003 at rank 64. These observations support the assumption on this model, task, rank range, and set of checkpoints; a finite sweep cannot establish uniform bounds for arbitrary models or unbounded ranks.

The role of the assumption can be separated from the scale algebra. For a common local unscaled gradient $\boldsymbol{g}_t(r)$ and residual-noise variance σ_ξ^2, the leading-order FD-SNR expressions under the

Gaussian direction convention of the analysis are

$$\mathsf{SNR}_{\mathrm{naive}}(r) \simeq \frac{2\alpha^2\mu^2\|\boldsymbol{g}_t(r)\|^2}{r^2\sigma_\xi^2}, \quad \mathsf{SNR}_{\mathrm{AR1\text{-}ZO}}(r) \simeq \frac{2\alpha^2\mu^2\|\boldsymbol{g}_t(r)\|^2}{\sigma_\xi^2}. \tag{8}$$

Thus pre-oracle calibration removes the explicit r^{-2} multiplier without Assumption C.3; characterizing the restored absolute FD-SNR as $\Theta(1)$ additionally requires its rank-uniform gradient bounds, with the other oracle parameters fixed. This is a local comparison of coefficients, not an assertion that differently scaled training trajectories have identical dense gradients or factor norms. Rank may still affect the unscaled gradient indirectly through the learned factors and trajectory.

B.5 Untuned Baselines and Downstream Exceptions

Table 8 adds untuned base-model scores under the same held-out protocol as the corresponding fine-tuned models. AR1-ZO improves all 24 model–task scores over the base model, by 2.0–60.8 points and 23.3 points on average. Relative to the direct full-adapter MeZO-LoRA control, it improves 23 of 24 scores by 4.2 points on average. These averages summarize the reported accuracy/F1 point differences across the six tasks; they are not a common statistical effect size. The base comparison establishes that the downstream gains reflect adaptation rather than small changes around an already strong untuned model.

Table 8: **Untuned and fine-tuned model performance.** BoolQ, CB, COPA, and WIC report accuracy; SQuAD and DROP report F1. All scores are on a 0–100 scale. ZO results use rank 64 and two forwards per step; full-model FT (Adam) is an FO reference. MeZO denotes the full-adapter LoRA query control throughout this table.

Model	Method	BoolQ	CB	COPA	WIC	SQuAD	DROP
OPT-2.7B	Base	50.0	46.4	69.0	53.0	28.2	11.0
	MeZO-LoRA	62.2	66.1	72.0	59.7	64.6	23.9
	AR1-ZO	64.6	69.6	79.0	60.3	77.8	27.8
	FT (Adam)	73.3	85.7	81.0	63.8	84.2	31.4
OPT-13B	Base	60.0	46.4	79.0	52.2	41.1	14.0
	MeZO-LoRA	64.8	66.1	84.0	54.9	78.7	30.1
	AR1-ZO	66.7	69.6	86.0	57.5	79.4	30.8
	FT (Adam)	76.9	84.1	79.0	70.1	84.9	31.3
Qwen3-1.7B	Base	65.4	41.1	70.0	50.0	17.5	7.9
	MeZO-LoRA	74.7	73.2	71.0	53.3	62.0	40.1
	AR1-ZO	78.0	87.5	72.0	56.6	76.7	42.9
Qwen3-32B	Base	73.8	44.6	78.0	61.8	19.8	11.4
	MeZO-LoRA	83.6	89.3	77.0	68.3	73.2	58.5
	AR1-ZO	85.6	90.4	88.0	71.5	80.6	52.2

Exceptions and interpretation. Qwen3-32B/DROP is the exception to the improvement over MeZO-LoRA (52.2 versus 58.5 F1). On the 12 OPT comparisons with LOZO in Table 1, AR1-ZO wins six, ties two, and loses four; the four deficits are 1.0–2.0 points. FO full fine-tuning is generally stronger because it uses backpropagated gradients and updates the full model. Different query geometries may provide task-dependent inductive biases, and finite-checkpoint accuracy or generation F1 can fluctuate even when training loss improves. These are plausible explanations for the exceptions, not causal conclusions established by the present experiments. The loss trajectories in Fig. 5 complement endpoint scores, but neither the trajectories nor the table imply a win for every task, seed, or checkpoint.

Backbone selection and scope. OPT supplies the complete prior-work baseline matrix because the compared methods provide public implementations and established protocols on that family. Dense Qwen3 models supply a focused transfer test with full-adapter MeZO-LoRA, naive atomwise querying, and AR1-ZO. This choice does not imply that LOZO or HIZOO are inapplicable to Qwen3. All evaluated Qwen3 variants, including the 8B memory-profile model below, are dense models. An

MoE extension would require specifying adapter placement in shared attention, routers, and expert-specific MLPs, while controlling token-dependent expert computation and the actual query budget; no MoE performance claim is made here.

B.6 Time to Target and Component-Level Memory

Matched time to an attainable loss. On OPT-2.7B/SQuAD, we compare full-model FT (Adam), MeZO-LoRA, and AR1-ZO with matched data, batch size, precision, and logging. The two ZO methods use rank 64 and two forward evaluations per step. We choose training loss 0.93, an attainable late-training level for MeZO-LoRA reached by all three methods. For each method, stable step time is the mean from a 500-step NVIDIA Jetson Thor profile after excluding the first 50 warm-up steps. Table 9 estimates time to target by multiplying the first-hit step by this mean. It therefore measures the optimization/steady-step trade-off and does not represent directly measured end-to-end elapsed time including initialization and evaluation.

Table 9: **Time to training loss 0.93 on OPT-2.7B/SQuAD.** Time to target is estimated from the first-hit step and stable mean step time. FO full fine-tuning is a reference with a different gradient-access contract.

Method	First-hit step	Seconds/step	Estimated seconds to target
FT (Adam)	140	0.9361	131.0
MeZO-LoRA	13,320	0.5420	7,219.7
AR1-ZO	690	0.5441	375.4

The two ZO methods have nearly identical step costs. AR1-ZO reaches the target with $19.3\times$ fewer steps and $19.2\times$ less estimated time than MeZO-LoRA. Its remaining gap to FO is $4.9\times$ in steps and $2.9\times$ in estimated time: backpropagation remains fastest in this comparison when feasible. The memory profiles below quantify the complementary resource benefit on a separate device and workload; their absolute numbers should not be conflated with the Thor timing profile.

Memory components and gradient checkpointing. We profile fixed $4{,}096$-token inputs on an NVIDIA H100 80GB, using two warm-up steps and eight measured steps. The Qwen3-1.7B profile uses batch size 2. Saved backward tensors are tracked with `saved_tensors_hooks`, deduplicating storage. Table 10 separates weights, adapters, gradients, Adam states, saved backward activations, and current-forward temporary/workspace memory, and reports peak allocated GPU memory in GiB. Entries are rounded independently.

Table 10: **Memory decomposition on Qwen3-1.7B.** H100 80GB, batch size 2, sequence length $4{,}096$. All values are GiB. GC denotes gradient checkpointing; "Saved act." denotes tensors retained for backward.

Method	Base weights	Adapter	Gradients	Adam states	Saved act.	Forward workspace	Peak allocated
Full FT (AdamW)	3.205	0	3.205	6.410	29.168	2.371	44.358
Full FT + GC	3.205	0	3.205	6.410	3.318	1.529	17.666
LoRA-AdamW	3.205	0.024	0.024	0.048	22.504	5.552	31.356
LoRA-AdamW + GC	3.205	0.024	0.024	0.048	3.255	4.710	12.295
AR1-ZO	3.205	0.024	0	0	0	7.862	11.091

LoRA restricts gradients and Adam states to the adapter, reducing persistent parameter-related memory from 12.819 to 3.301 GiB. Gradient checkpointing further reduces saved activations while retaining backward computation, gradients, and Adam states. AR1-ZO uses inference-mode forwards and eliminates all three of these backward-related components. Its peak is 75.0% below full FT, 37.2% below full FT with GC, 64.6% below LoRA-AdamW, and 9.8% below LoRA-AdamW with GC. The remaining 7.862 GiB of temporary/workspace memory is dominated by current-forward logits and kernels, so removing saved backward activations does not eliminate forward memory requirements.

At Qwen3-8B, AR1-ZO reduces peak memory by 57.8% relative to LoRA-AdamW and 30.1% relative to LoRA-AdamW with GC (Table 11). Its 18.229 GiB comprises 15.256 GiB of base weights,

Table 11: **Memory at a larger model scale.** Qwen3-8B on H100 80GB with batch size 1, sequence length 4,096, and the same warm-up and measurement protocol. Values are GiB.

Method	Saved backward activations	Peak allocated
LoRA-AdamW	25.408	43.219
LoRA-AdamW + GC	2.346	26.086
AR1-ZO	0	18.229

a 0.057 GiB adapter, and 2.916 GiB of forward workspace. Weights and forward workspace still grow with model size and batch/sequence dimensions. The absence of saved backward activations, gradients, and optimizer states is preserved by the forward-only update, but these measurements do not imply a hardware-independent memory ratio.

B.7 DoRA and the Scope of Factorized Adaptation

DoRA [13] augments low-rank direction factors $\boldsymbol{A}, \boldsymbol{B}$ with independent magnitude parameters. In the native DoRA extension, AR1-ZO perturbs one matched direction atom $(\boldsymbol{b}_k, \boldsymbol{a}_k)$ together with the magnitude parameters, holding inactive direction atoms fixed. The magnitude coordinates add to the active query dimension, so the pure-LoRA dimension q does not describe the entire DoRA query. Both ZO variants still use two forward evaluations per step. Table 12 reports Qwen3-1.7B/BoolQ at rank 64 with matched data and precision. The final accuracy uses the stated training budget; the common training-loss threshold 0.50 lies in MeZO+DoRA's late-training convergence region and is attained by all three methods.

Table 12: **Native DoRA extension.** Qwen3-1.7B/BoolQ, rank 64. FO DoRA uses AdamW. The first-hit step and time refer to training loss 0.50; accuracy is measured after the full training budget.

Method	Budget (steps)	Accuracy (%)	First-hit step	Time to target (s)
FO DoRA + AdamW	2,000	83.8	930	1,166
MeZO + DoRA	20,000	75.0	6,640	7,810
AR1-ZO + DoRA	20,000	77.2	2,320	4,619

AR1-ZO+DoRA gains 2.2 accuracy points over MeZO+DoRA and reaches the target with 65.1% fewer steps and 40.9% less time. FO remains faster and more accurate. The improvement supports transfer of the query principle to a direction–magnitude parameterization, but equal query counts alone do not imply equal per-step wall time across implementations.

Expressivity versus practical full-model training. Any matrix update $\Delta\boldsymbol{W}$ admits a factorization $\boldsymbol{BA}$ when $r \geq \operatorname{rank}(\Delta\boldsymbol{W})$. Factorizing every linear weight can therefore recover arbitrary matrix-update expressivity, while biases and normalization scales require separate ZO blocks. This is an algebraic observation, not a convergence or memory guarantee for full-model factorized training. The experiments validate low-rank LoRA and the stated DoRA extension; they do not validate arbitrary full-rank factorizations or every PEFT variant. Low rank can nevertheless approach the full-model FO reference on individual tasks, such as OPT-2.7B/COPA (79.0 versus 81.0) and DROP (27.8 versus 31.4).

B.8 Open-Ended Generation and Unseen-Task Instruction Following

The original suite spans reading comprehension (BoolQ), natural-language inference (CB), causal reasoning (COPA), lexical disambiguation (WIC), and generated question answers (SQuAD and DROP, with discrete/numerical reasoning in DROP). We extend this suite with abstractive dialogue summarization and instruction-conditioned generalization to unseen tasks, which exercise capabilities beyond labels and short answers.

SAMSum dialogue summarization. SAMSum [6] pairs dialogues with abstractive summaries. The prompt contains the instruction "Summarize the following conversation" followed by a multi-turn dialogue, and the model generates a free-form summary. We evaluate Qwen3-1.7B on all 819

test examples. All fine-tuning methods use rank 64, batch size 8, and greedy decoding with at most 128 new tokens. MeZO-LoRA and AR1-ZO each use 20,000 steps and two forwards per step; Adam-LoRA uses 2,000 steps as an FO reference. Table 13 reports the complete-test-set results.

Table 13: **SAMSum abstractive summarization.** Qwen3-1.7B on the complete 819-example test set. ROUGE scores are on a 0–100 scale.

Method	ROUGE-1	ROUGE-2	ROUGE-L
Base model	34.50	12.43	25.87
MeZO-LoRA	39.73	15.93	31.17
AR1-ZO	39.85	16.27	32.60
Adam-LoRA (FO reference)	47.46	22.90	39.24

AR1-ZO improves over the base model by 5.35 ROUGE-1, 3.84 ROUGE-2, and 6.73 ROUGE-L, and exceeds matched-budget MeZO-LoRA by 1.43 ROUGE-L. The FO reference remains stronger. Table 14 illustrates improved content selection in a separate Qwen3-0.6B development run: the trained output includes the total price and omits unsupported or unnecessary details. This example is qualitative evidence from a different model size and is not part of the Qwen3-1.7B aggregate comparison.

Table 14: **Qualitative SAMSum example from Qwen3-0.6B.** A separate development run, distinct from Table 13. The dialogue is abridged; the reference and generated summaries are shown in full.

Instruction: Summarize the following conversation.

Dialogue (abridged):
Jessica: I went to the second hand shop downtown.
Jessica: I bought a table, six chairs, a vase, and a pile of clothes.
Raphael: Is the table wood?
Jessica: Yes, oak.
Frank: I love the black dress.
Jessica: I'm so happy with what I bought.
Jessica: And the best is I paid only 70 euros for all of that!

Reference: Jessica bought a table, six chairs, a vase and a pile of clothes at the second-hand shop downtown. She paid 70 euros for everything.

Before AR1-ZO: Jessica bought a table, six chairs, a vase, and a pile of clothes at a second-hand shop downtown. She received photos and described the items, mentioning the table is wood (oak), the clothes are black, and she was happy with the purchase.
ROUGE-1/2/L: 58.82 / 48.48 / 52.94

After AR1-ZO: Jessica bought a table, six chairs, a vase, and a pile of clothes at a second-hand shop downtown. She paid 70 euros for everything.
ROUGE-1/2/L: 92.00 / 87.50 / 92.00

Super-NaturalInstructions cross-task evaluation. Super-NaturalInstructions [22] evaluates generalization from natural-language task instructions to unseen tasks. Our fixed split uses 500 official training tasks with two unique instances per task (1,000 examples), 50 disjoint official training tasks for validation with four instances each (200 examples), and all 119 official unseen test tasks with 20 unique instances each (2,380 examples). Selection uses seed 42. Duplicate entries are grouped by instance ID; training selects one target deterministically while evaluation retains all valid references. At test time, Qwen3-1.7B receives only the unseen task's definition and input in a user message, with thinking disabled, and generates one answer by greedy decoding with at most 128 new tokens. Each answer is scored against all references, retaining the maximum score for each metric. Scores are averaged first within each task and then across the 119 tasks to obtain task-macro ROUGE-L and exact match.

Both ZO methods use rank 64, 20,000 steps, batch size 16, maximum training sequence length 1,024, and the same two-forward query budget. Cross-entropy is computed only on target and end-of-sequence tokens. Table 15 shows gains over both the untuned model and MeZO-LoRA. From the unrounded scores, AR1-ZO improves over the base model by 17.93 ROUGE-L and 20.92 exact-match points, and over MeZO-LoRA by 0.95 and 1.39 points, respectively. Because test tasks are

officially held out and absent from training, this protocol measures transfer to unseen tasks rather than an instance split within training tasks. Together with SAMSum, it extends the empirical scope to free-form generation and instruction-conditioned cross-task generalization.

Table 15: **Super-NaturalInstructions on** 119 **unseen tasks.** Qwen3-1.7B; task-macro scores on a 0–100 scale. Differences in the text are computed from unrounded scores.

Method	Macro ROUGE-L	Macro exact match
Base model	27.46	5.80
MeZO-LoRA	44.44	25.34
AR1-ZO	45.40	26.72

C Theoretical Proofs

C.1 Preliminaries and Assumptions

Before presenting the proofs, we formalize the necessary mathematical assumptions and standard properties of the zeroth-order oracle.

Assumption C.1 (L-smoothness)**.** The objective function $\mathcal{L}(\theta)$ is differentiable and its gradient is L-Lipschitz. For any θ, θ', there exists a constant $L > 0$ such that:

$$\|\nabla\mathcal{L}(\theta') - \nabla\mathcal{L}(\theta)\| \le L\|\theta' - \theta\|. \tag{9}$$

This yields the standard quadratic upper bound:

$$\mathcal{L}(\theta') \le \mathcal{L}(\theta) + \langle\nabla\mathcal{L}(\theta),\, \theta' - \theta\rangle + \frac{L}{2}\|\theta' - \theta\|^2. \tag{10}$$

Assumption C.2 (Bounded Evaluation Noise)**.** In practice, we only have access to a stochastic realization of the loss function. Paired evaluations may share the same mini-batch or other common random numbers; after conditioning on this shared information, we model the remaining oracle error as $\tilde{\mathcal{L}}(\theta) = \mathcal{L}(\theta) + \xi$, where ξ is an independent zero-mean residual satisfying:

$$\mathbb{E}[\xi] = 0, \qquad \mathbb{E}\big[\xi^2\big] = \sigma_\xi^2. \tag{11}$$

The residual errors in two evaluations $\tilde{\mathcal{L}}(\theta + \mu\boldsymbol{z})$ and $\tilde{\mathcal{L}}(\theta - \mu\boldsymbol{z})$ are modeled as independent realizations ξ_+ and ξ_-.

Assumption C.3 (Active-atom gradient regularity)**.** Along the analyzed iterates, the unscaled active factor-coordinate gradient

$$\boldsymbol{g}_t \triangleq \begin{bmatrix} \boldsymbol{G}\,\boldsymbol{a}_{k(t)} \\ \boldsymbol{G}^\top \boldsymbol{b}_{k(t)} \end{bmatrix} \in \mathbb{R}^q, \qquad q = d_{\text{out}} + d_{\text{in}},$$

satisfies $\|\boldsymbol{g}_t\| = \Theta(1)$ in the LoRA rank r; that is, $\|\boldsymbol{g}_t\|$ admits upper and lower bounds that depend only on the dense gradient $\boldsymbol{G}$ and on the smoothness/Lipschitz constants of $\mathcal{L}$, not on r. We further assume the queried atom has bounded factor norms, $\|\boldsymbol{b}_{k(t)}\|, \|\boldsymbol{a}_{k(t)}\| = \Theta(1)$, the standard local regularity used to interpret the active alignment β in Theorem 4.7. Both conditions are *local* statements about the queried atom $k(t)$ and do not constrain the magnitudes of the remaining $r-1$ stored atoms. The upper-bound half $\|\boldsymbol{g}_t\| = \mathcal{O}(1)$ alone suffices for the convergence bound in Theorem 4.4 Part II; the full $\Theta(1)$ regularity is used by Part I to make SNR_t rank-invariant in a precise rate sense.

The rank-uniform bounds in Assumption C.3 are local regularity conditions, not a claim of identical norms across ranks. Appendix B.4 reports direct finite-rank measurements and distinguishes removal of the explicit r^{-2} FD-SNR penalty from the additional assumption needed for absolute $\Theta(1)$ FD-SNR.

Assumption C.4 (Stable Represented Adapter Scale)**.** The represented adapter remains in a bounded forward-scale regime along the analyzed iterates:

$$\left\| \frac{\gamma}{r} \sum_{k=1}^{r} \boldsymbol{b}_k \boldsymbol{a}_k^\top \right\|_F = \mathcal{O}(1).$$

For AR1-ZO, where $\gamma/r = \alpha$, this is equivalent to $\left\|\sum_{k=1}^{r} \boldsymbol{b}_k \boldsymbol{a}_k^\top\right\|_F = \mathcal{O}(1)$. The assumption rules out interpreting the scaling law as an uncontrolled increase of the entire adapter output; the active FD-SNR algebra itself only depends on the active coefficient γ/r.

Relation to Assumption C.3. Assumption C.4 concerns the represented adapter $\|\Delta \boldsymbol{W}\|_F$, a sum over all r stored atoms, while Assumption C.3 concerns the queried atom only (its active gradient norm and its factor norms). The two statements constrain different physical objects—one global summed, one local active—and are used in disjoint segments of the proofs. In particular, with per-atom factor norms $\Theta(1)$ on the queried atom, generic incoherent configurations of the remaining stored atoms would give $\left\|\sum_k \boldsymbol{b}_k \boldsymbol{a}_k^\top\right\|_F = \Theta(\sqrt{r})$, so Assumption C.4 is *not* an incoherence condition: it is a spectrum-control condition on the trained adapter, requiring $\boldsymbol{BA}$ to remain in a regime where a small number of singular components dominate and the Frobenius norm of the sum stays $\mathcal{O}(1)$ in r. This regime is observed empirically in trained LoRA, where the effective rank of $\boldsymbol{BA}$ is well below the nominal r; it is a property of the optimization trajectory rather than a geometric prior on the atoms, and it imposes no constraint on the active-atom regularity asserted by Assumption C.3.

C.2 Auxiliary Lemmas

Lemma C.5 (Active-Atom Gradient under Factored Parameterization)**.** *Consider the factored update $\Delta \boldsymbol{W} = \frac{\gamma}{r}\sum_{k=1}^{r} \boldsymbol{b}_k \boldsymbol{a}_k^\top$, where $\boldsymbol{G} = \nabla_{\boldsymbol{W}} \mathcal{L} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ is the dense matrix gradient. The active-atom gradient with respect to atom k is, writing $\boldsymbol{b}$ and $\boldsymbol{a}$ in gradient subscripts for the active factors $\boldsymbol{b}_k$ and $\boldsymbol{a}_k$:*

$$\nabla_{\boldsymbol{b}} \mathcal{L} = \frac{\gamma}{r} \boldsymbol{G} \boldsymbol{a}_k, \qquad \nabla_{\boldsymbol{a}} \mathcal{L} = \frac{\gamma}{r} \boldsymbol{G}^\top \boldsymbol{b}_k. \tag{12}$$

Proof. By the chain rule, since $\boldsymbol{W} = \boldsymbol{W}_0 + \frac{\gamma}{r}\sum_j \boldsymbol{b}_j \boldsymbol{a}_j^\top$:

$$\frac{\partial \boldsymbol{W}}{\partial [\boldsymbol{b}]_i} = \frac{\gamma}{r} \boldsymbol{e}_i \boldsymbol{a}_k^\top.$$

Therefore:

$$[\nabla_{\boldsymbol{b}} \mathcal{L}]_i = \operatorname{tr}\Big(\boldsymbol{G}^\top \cdot \frac{\gamma}{r} \boldsymbol{e}_i \boldsymbol{a}_k^\top\Big) = \frac{\gamma}{r} \boldsymbol{e}_i^\top \boldsymbol{G} \boldsymbol{a}_k = \frac{\gamma}{r} [\boldsymbol{G} \boldsymbol{a}_k]_i.$$

Stacking over all i yields $\nabla_{\boldsymbol{b}} \mathcal{L} = \frac{\gamma}{r} \boldsymbol{G} \boldsymbol{a}_k$. The derivation for $\nabla_{\boldsymbol{a}} \mathcal{L}$ is analogous. □

Lemma C.6 (Projected ZO-RGE moment bounds)**.** *Let d denote the full LoRA coordinate dimension for the adapted matrix, and let $\boldsymbol{P} \in \mathbb{R}^{d \times q}$ be the orthogonal basis spanning atom k's active coordinates, with $\boldsymbol{P}^\top \boldsymbol{P} = \boldsymbol{I}_q$ and $q = d_{\text{in}} + d_{\text{out}}$, and let $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}_q)$. The two-point Randomized Gradient Estimator (RGE):*

$$\hat{\boldsymbol{g}} = \frac{\tilde{\mathcal{L}}(\theta + \mu \boldsymbol{P} \boldsymbol{z}) - \tilde{\mathcal{L}}(\theta - \mu \boldsymbol{P} \boldsymbol{z})}{2\mu} \boldsymbol{z}$$

satisfies the following standard projected two-point moment bounds under the smoothness conditions needed for the central finite-difference remainder:

1. ***Bias Bound:*** $\|\mathbb{E}_{\boldsymbol{z},\xi}[\hat{\boldsymbol{g}}] - \nabla_k \mathcal{L}\| = \mathcal{O}(\mu^2 L\, q)$.
2. ***Second Moment Bound:*** $\mathbb{E}_{\boldsymbol{z},\xi}\big[\|\hat{\boldsymbol{g}}\|^2\big] \le (q+3)\,\|\nabla_k \mathcal{L}\|^2 + C_1 \mu^2 L^2 q^3 + \frac{C_2\, q\, \sigma_\xi^2}{2\mu^2}$, *for universal constants $C_1, C_2 > 0$. The leading coefficient $(q+3)$ is a clean upper bound that absorbs signal–bias cross terms arising from the $\boldsymbol{z}$-dependent Hessian remainder; the exact signal-only second moment is $(q+2)\|\nabla_k \mathcal{L}\|^2$.*

Proof of Lemma C.6 follows from standard zeroth-order optimization literature [15, 12].

C.3 Proof of Lemma 4.1: Complete Rank-1 Atom Blocks

Proof. For the atom map $\phi_k(\boldsymbol{b}_k, \boldsymbol{a}_k) = \boldsymbol{b}_k \boldsymbol{a}_k^\top$, the Fréchet derivative in direction $(\delta \boldsymbol{b}, \delta \boldsymbol{a})$ is

$$D\phi_k(\delta \boldsymbol{b}, \delta \boldsymbol{a}) = \frac{d}{d\epsilon}(\boldsymbol{b}_k + \epsilon \delta \boldsymbol{b})(\boldsymbol{a}_k + \epsilon \delta \boldsymbol{a})^\top \Big|_{\epsilon=0} = \delta \boldsymbol{b}\, \boldsymbol{a}_k^\top + \boldsymbol{b}_k\, \delta \boldsymbol{a}^\top.$$

If only $\boldsymbol{b}_k$ is perturbed, the attainable tangent matrices have the form $\delta \boldsymbol{b}\, \boldsymbol{a}_k^\top$ and keep the right factor fixed at $\boldsymbol{a}_k$. If only $\boldsymbol{a}_k$ is perturbed, they have the form $\boldsymbol{b}_k\, \delta \boldsymbol{a}^\top$ and keep the left factor fixed at $\boldsymbol{b}_k$. When both factors are nonzero, each one-sided family is a strict subset of the full factor-coordinate

tangent generated by $D\phi_k(\delta \boldsymbol{b}, \delta \boldsymbol{a})$. Therefore a complete atom-local LoRA-native factor-coordinate perturbation includes the matched pair $(\boldsymbol{b}_k, \boldsymbol{a}_k)$.

The stored coordinates of this matched pair have dimension $d_{\text{out}} + d_{\text{in}}$. The rank-1 matrix manifold itself has one scaling gauge redundancy, since $(\boldsymbol{b}, \boldsymbol{a})$ and $(c\boldsymbol{b}, \boldsymbol{a}/c)$ represent the same matrix for $c \neq 0$, so its intrinsic dimension is $d_{\text{out}} + d_{\text{in}} - 1$. AR1-ZO, however, queries the stored LoRA factors through a black-box oracle, not a quotient manifold coordinate chart; the factor-coordinate query therefore perturbs $d_{\text{out}} + d_{\text{in}}$ variables. A complete block of m matched atoms is the direct product of m such blocks and has query dimension $m(d_{\text{out}} + d_{\text{in}})$. □

C.4 Proof of Proposition 4.2: Active-Atom ZO Dimension

Proof. Fix iteration t and write $k = k(t)$. Let $\boldsymbol{P}_t$ inject an active perturbation $\boldsymbol{z} \in \mathbb{R}^q$ into the factor coordinates of atom k, and define the proof-local active-block Gaussian smoothing $\mathcal{L}_{\mu,k(t)}(\theta) = \mathbb{E}_{\boldsymbol{z}}[\mathcal{L}(\theta + \mu \boldsymbol{P}_t \boldsymbol{z})]$. The standard Gaussian smoothing identity for the two-point randomized gradient estimator gives

$$\mathbb{E}_{\boldsymbol{z},\xi}\left[\frac{\tilde{\mathcal{L}}(\theta + \mu \boldsymbol{P}_t \boldsymbol{z}) - \tilde{\mathcal{L}}(\theta - \mu \boldsymbol{P}_t \boldsymbol{z})}{2\mu} \boldsymbol{z}\right] = \nabla_{k(t)} \mathcal{L}_{\mu,k(t)}(\theta),$$

because the residual oracle noise is zero mean after conditioning on the paired query protocol.

Lemma C.6 applies with active dimension $q = d_{\text{out}} + d_{\text{in}}$. Combining its second-moment and bias bounds gives the usual mean-squared error form

$$\mathbb{E}\Big[\|\hat{\nabla}_{k(t)} \mathcal{L} - \nabla_{k(t)} \mathcal{L}(\theta)\|^2\Big] \lesssim q \|\nabla_{k(t)} \mathcal{L}(\theta)\|^2 + L^2 \mu^2 q^3 + \frac{\sigma_\xi^2 q}{\mu^2},$$

up to universal constants. The dimension appearing in every term is the dimension of the sampled vector $\boldsymbol{z}$, namely q. If instead one perturbs all r LoRA atoms of the same matrix at once, the same standard ZO bound is obtained with q replaced by rq. □

C.5 Proof of Theorem 4.3: Directional Collapse under Naive Scaling

Theorem 4.3 (Restated). *Under Assumptions C.1, C.2, C.3 and naive $\gamma = \alpha$, the active signal is $\Theta(r^{-1})$ and active FD-SNR $\mathsf{SNR}_t = \Theta(r^{-2})$. With $r_c = \Theta(\alpha\mu\|\boldsymbol{g}_t\|/\sigma_\xi) = \Theta(\alpha\mu/\sigma_\xi)$, $r \gg r_c$ makes the expected cosine with the active gradient vanish.*

Proof. Let θ denote the full parameter vector. At step t, we freeze all parameters except atom k, the coordinate pair $(\boldsymbol{b}_k, \boldsymbol{a}_k)$ with dimension $q = d_{\text{in}} + d_{\text{out}}$.

Step 1: Signal Magnitude Decay.

Under the naive scaling $\gamma = \alpha$, Lemma C.5 gives:

$$\nabla_{\boldsymbol{b}} \mathcal{L} = \frac{\alpha}{r} \boldsymbol{G} \boldsymbol{a}_k, \qquad \nabla_{\boldsymbol{a}} \mathcal{L} = \frac{\alpha}{r} \boldsymbol{G}^\top \boldsymbol{b}_k. \tag{13}$$

Define the unscaled active factor-coordinate gradient:

$$\boldsymbol{g}_k \;=\; \begin{bmatrix} \boldsymbol{G}\boldsymbol{a}_k \\ \boldsymbol{G}^\top \boldsymbol{b}_k \end{bmatrix} \;\in\; \mathbb{R}^q. \tag{14}$$

Then at rank r:

$$\nabla_k \mathcal{L} = \frac{\alpha}{r} \boldsymbol{g}_k, \qquad \|\nabla_k \mathcal{L}\| = \frac{\alpha}{r} \|\boldsymbol{g}_k\|. \tag{15}$$

By Assumption C.3, $\|\boldsymbol{g}_k\| = \Theta(1)$ in r, so the signal magnitude decays as $\Theta(1/r)$ rather than merely as $\mathcal{O}(1/r) \cdot \|\boldsymbol{g}_k\|$ with a possibly r-dependent prefactor. □

Step 2: Finite-Difference Signal–Noise Decomposition.

The two-point FD observation along $\boldsymbol{z} \sim \mathcal{N}(0, \boldsymbol{I}_q)$ is decomposed via Taylor expansion:

$$\Delta_\mu \;=\; \frac{\tilde{\mathcal{L}}(\theta + \mu \boldsymbol{P}\boldsymbol{z}) - \tilde{\mathcal{L}}(\theta - \mu \boldsymbol{P}\boldsymbol{z})}{2\mu} \;=\; \underbrace{\boldsymbol{z}^\top \nabla_k \mathcal{L}}_{\text{signal}} + \underbrace{\mathcal{O}(\mu^2)}_{\text{bias}^2} + \underbrace{\frac{\xi_+ - \xi_-}{2\mu}}_{\text{noise } \zeta}. \tag{16}$$

The full ZO estimator is then $\hat{\boldsymbol{g}} = \Delta_\mu \cdot \boldsymbol{z} = (\boldsymbol{z}^\top \nabla_k \mathcal{L})\, \boldsymbol{z} + \mathcal{O}(\mu^2)\, \boldsymbol{z} + \zeta\, \boldsymbol{z}$.

Signal power:

$$\mathbb{E}\big[(\boldsymbol{z}^\top \nabla_k \mathcal{L})^2\big] = \|\nabla_k \mathcal{L}\|^2 = \frac{\alpha^2 \|\boldsymbol{g}_k\|^2}{r^2}. \tag{17}$$

Residual noise power (independent of r):

$$\mathbb{E}\big[\zeta^2\big] = \frac{\sigma_\xi^2}{2\mu^2}. \tag{18}$$

□

Step 3: Active FD-SNR and Critical Rank.

We define the active FD-SNR as the ratio of signal power to residual noise power:

$$\mathsf{SNR}(r) = \frac{\alpha^2 \|\boldsymbol{g}_k\|^2 / r^2}{\sigma_\xi^2/(2\mu^2)} = \frac{2\alpha^2 \mu^2 \|\boldsymbol{g}_k\|^2}{r^2 \sigma_\xi^2}. \tag{19}$$

This decays as $\mathcal{O}(1/r^2)$. Setting $\mathsf{SNR} = 1$ and solving for r:

$$r_c = \frac{\sqrt{2}\, \alpha \mu \|\boldsymbol{g}_k\|}{\sigma_\xi} = \Theta\!\left(\frac{\alpha \mu \|\boldsymbol{g}_k\|}{\sigma_\xi}\right). \tag{20}$$

Under Assumption C.3, $\|\boldsymbol{g}_k\| = \Theta(1)$, so $r_c = \Theta(\alpha\mu/\sigma_\xi)$ depends only on the smoothing radius, the active LoRA scale, and the residual evaluation noise. □

Step 4: Cosine Similarity Collapse.

From the decomposition in (16), the ZO estimator is $\hat{\boldsymbol{g}} = (\boldsymbol{z}^\top \boldsymbol{g})\, \boldsymbol{z} + \zeta\, \boldsymbol{z}$ (dropping the $\mathcal{O}(\mu^2)$ bias under the standard assumption that the smoothing parameter μ is chosen sufficiently small, e.g., $\mu = \mathcal{O}(q^{-1/2})$, such that the finite-difference bias is strictly dominated by the stochastic noise ζ), where $\boldsymbol{g} = \nabla_k \mathcal{L}$.

We analyze $\mathbb{E}[\cos(\hat{\boldsymbol{g}},\, \boldsymbol{g})]$ by computing the numerator and denominator separately.

Numerator:

$$\mathbb{E}\big[\hat{\boldsymbol{g}}^\top \boldsymbol{g}\big] = \mathbb{E}\big[(\boldsymbol{z}^\top \boldsymbol{g})^2\big] + \mathbb{E}\big[\zeta\, (\boldsymbol{z}^\top \boldsymbol{g})\big] = \|\boldsymbol{g}\|^2. \tag{21}$$

Here, $\mathbb{E}\big[\zeta\, (\boldsymbol{z}^\top \boldsymbol{g})\big] = 0$ since ζ and $\boldsymbol{z}$ are independent.

Denominator: Since the bias has been dropped, the second moment of the bias-free estimator $\hat{\boldsymbol{g}} = (\boldsymbol{z}^\top \boldsymbol{g})\boldsymbol{z} + \zeta \boldsymbol{z}$ uses the exact signal coefficient $(q+2)$, and the smoothness remainder $C_1 \mu^2 L^2 q^3$ from Lemma C.6 is likewise negligible under $\mu = \mathcal{O}(q^{-1/2})$:

$$\mathbb{E}\big[\|\hat{\boldsymbol{g}}\|^2\big] = (q+2)\|\boldsymbol{g}\|^2 + \frac{\sigma_\xi^2 q}{2\mu^2}. \tag{22}$$

By the concentration of measure in high dimensions and standard zeroth-order bounds (cf. [15], Lemma 3), the random variable $\|\hat{\boldsymbol{g}}\|^2$ tightly concentrates around its expectation. Thus, the expected cosine similarity can be approximated as:

$$\begin{aligned} \mathbb{E}[\cos(\hat{\boldsymbol{g}},\, \boldsymbol{g})] \;&\asymp\; \frac{\|\boldsymbol{g}\|^2}{\|\boldsymbol{g}\| \sqrt{(q+2)\,\|\boldsymbol{g}\|^2 + \sigma_\xi^2\, q/(2\mu^2)}} \\ &=\; \frac{\|\boldsymbol{g}\|}{\sqrt{(q+2)\,\|\boldsymbol{g}\|^2 + \sigma_\xi^2\, q/(2\mu^2)}}. \end{aligned} \tag{23}$$

Substituting $\|\boldsymbol{g}\| = \alpha \|\boldsymbol{g}_k\|/r$:

$$\mathbb{E}[\cos(\hat{\boldsymbol{g}}_{\text{naive}},\, \boldsymbol{g})] \;\asymp\; \frac{\alpha \|\boldsymbol{g}_k\|/r}{\sqrt{(q+2)\,\alpha^2 \|\boldsymbol{g}_k\|^2/r^2 + \sigma_\xi^2\, q/(2\mu^2)}}. \tag{24}$$

[2]The $\mathcal{O}(\mu^2)$ bias of central differences assumes standard higher-order smoothness (bounded third derivatives). Under only L-smoothness (Assumption C.1), the bias is $\mathcal{O}(\mu)$, which is still strictly dominated by the stochastic noise ζ for small μ.

When $r \gg r_c$, the residual noise term dominates the denominator:

$$\frac{\sigma_\xi^2\, q}{2\mu^2} \;\gg\; \frac{(q+2)\,\alpha^2 \|\boldsymbol{g}_k\|^2}{r^2}, \tag{25}$$

so the denominator $\approx \sqrt{\sigma_\xi^2\, q/(2\mu^2)}$, and:

$$\mathbb{E}[\cos(\hat{\boldsymbol{g}}_{\text{naive}},\, \boldsymbol{g})] \;\approx\; \frac{\alpha\|\boldsymbol{g}_k\|/r}{\sigma_\xi\,\sqrt{q/(2\mu^2)}} \;=\; \frac{\sqrt{2}\,\alpha\mu\,\|\boldsymbol{g}_k\|}{r\,\sigma_\xi\,\sqrt{q}} \;=\; \mathcal{O}\!\left(\frac{r_c}{r\,\sqrt{q}}\right) \;\to\; 0. \tag{26}$$

This establishes **directional collapse**: the ZO estimate loses all correlation with the true gradient direction. The inverse dependence on both r and $\sqrt{q}$ implies that in high-dimensional embedding spaces, even moderate ranks can trigger collapse when σ_ξ is non-vanishing. □

Remark C.7. This collapse is qualitatively different from slow convergence. In first-order optimization, scaling the gradient by $1/r$ merely reduces step size while preserving direction. In ZO, once the signal falls below the noise floor, the *direction itself* is destroyed. No learning rate adjustment can repair a noise-dominated direction estimate.

C.6 Proof of Theorem 4.4: Signal Restoration and Active-Atom Stationarity

Theorem 4.4 (Restated). ***(Part I, FD-SNR.)*** *Under Assumptions C.1, C.2, C.3 and topology-aware scaling $\gamma = \alpha r$, the active FD-SNR satisfies*

$$\mathsf{SNR}_t \asymp \frac{\alpha^2\mu^2\|\boldsymbol{g}_t\|^2}{\sigma_\xi^2} = \Theta(1) \;\; \textit{in } r.$$

(Part II, Convergence.) *Under Assumptions C.1, C.2, C.4, the upper-bound half of Assumption C.3 (i.e., $\|\boldsymbol{g}_t\| = \mathcal{O}(1)$ along iterates, with bounded active factor norms $\|\boldsymbol{b}_{k(t)}\|, \|\boldsymbol{a}_{k(t)}\| = \mathcal{O}(1)$), and the projected RGE moment bounds of Lemma C.6, with topology-aware scaling $\gamma = \alpha r$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla_{k(t)}\mathcal{L}(\theta_t)\|^2\big] \le \mathcal{O}\!\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}\!\left(\mu^4 L^2 q^2 + \frac{\sigma_\xi^2\, q}{T^{1/2}\mu^2}\right). \tag{27}$$

Part I uses Assumption C.3 to make the active FD-SNR a true rank-invariant quantity; Part II uses Assumption C.4 to rule out trivial inflation of the represented adapter under $\gamma = \alpha r$, and depends on Assumption C.3 only through the upper bound $\|\boldsymbol{g}_t\| = \mathcal{O}(1)$.

Proof. **Part I: Signal Restoration**

Step 1: Active-atom gradient under $\gamma = \alpha r$.

By Lemma C.5, the active-atom gradient with respect to atom k under general scaling γ is:

$$\nabla_{\boldsymbol{b}}\mathcal{L} = \frac{\gamma}{r}\,\boldsymbol{G}\boldsymbol{a}_k, \qquad \nabla_{\boldsymbol{a}}\mathcal{L} = \frac{\gamma}{r}\,\boldsymbol{G}^\top\boldsymbol{b}_k. \tag{28}$$

Substituting $\gamma = \alpha r$:

$$\nabla_{\boldsymbol{b}}\mathcal{L} = \frac{\alpha r}{r}\,\boldsymbol{G}\boldsymbol{a}_k = \alpha\,\boldsymbol{G}\boldsymbol{a}_k, \qquad \nabla_{\boldsymbol{a}}\mathcal{L} = \alpha\,\boldsymbol{G}^\top\boldsymbol{b}_k. \tag{29}$$

The factor r in γ exactly cancels the $1/r$ from the factored parameterization. The concatenated active-atom gradient is therefore:

$$\nabla_k\mathcal{L} = \alpha\begin{bmatrix}\boldsymbol{G}\boldsymbol{a}_k \\ \boldsymbol{G}^\top\boldsymbol{b}_k\end{bmatrix} = \alpha\boldsymbol{g}_k, \tag{30}$$

In particular, $\|\nabla_k\mathcal{L}\| = \alpha\|\boldsymbol{g}_k\|$, which is independent of the total rank r. □

Step 2: Active FD-SNR independence of r.

From Step 2 of the naive-scaling proof above, the active FD-SNR is:

$$\mathsf{SNR} = \frac{\text{signal power}}{\text{noise power}} = \frac{\|\nabla_k \mathcal{L}\|^2}{\sigma_\xi^2/(2\mu^2)} = \frac{2\alpha^2\mu^2\|\boldsymbol{g}_k\|^2}{\sigma_\xi^2}, \tag{31}$$

which is constant in r. By Assumption C.3, $\|\boldsymbol{g}_t\| = \Theta(1)$, hence $\mathsf{SNR}_t = \Theta(1)$ in r. Signal restoration is *measurement-level* and is not contingent on the represented adapter scale. □

Part II: Convergence

The convergence bound below uses Assumption C.4 to rule out the trivial reading that $\gamma = \alpha r$ merely inflates the represented adapter; it does *not* require the lower-bound half of Assumption C.3, since the descent inequality and the projected RGE moment bounds depend only on the active dimension q, on Lipschitz/smoothness constants, and on the upper bound $\|\boldsymbol{g}_t\| = \mathcal{O}(1)$.

Step 3: Per-step descent (Descent Lemma).

At step t, the active atom is $k(t)$ and the update is $\theta_{t+1} = \theta_t - \eta_t\,\hat{\boldsymbol{g}}_t$, where $\hat{\boldsymbol{g}}_t$ is the projected ZO estimator restricted to the active atom $k(t)$ as defined in Lemma C.6.

By Assumption C.1 (L-smoothness), the standard descent inequality applied to the full gradient gives $\mathcal{L}(\theta_{t+1}) \le \mathcal{L}(\theta_t) - \eta_t\,\langle \nabla\mathcal{L}(\theta_t),\, \hat{\boldsymbol{g}}_t\rangle + \frac{\eta_t^2 L}{2}\,\|\hat{\boldsymbol{g}}_t\|^2$. Since $\hat{\boldsymbol{g}}_t$ is supported entirely on the active atom $k(t)$ (all other atom coordinates are zero), the inner product reduces to $\langle \nabla\mathcal{L}(\theta_t),\, \hat{\boldsymbol{g}}_t\rangle = \langle \nabla_{k(t)}\mathcal{L}(\theta_t),\, \hat{\boldsymbol{g}}_t\rangle$. Thus:

$$\mathcal{L}(\theta_{t+1}) \le \mathcal{L}(\theta_t) - \eta_t\,\langle \nabla_{k(t)}\mathcal{L}(\theta_t),\, \hat{\boldsymbol{g}}_t\rangle + \frac{\eta_t^2 L}{2}\,\|\hat{\boldsymbol{g}}_t\|^2. \tag{32}$$

Taking expectations over $\boldsymbol{z}_t$ and ξ_t conditioned on θ_t:

$$\mathbb{E}[\mathcal{L}(\theta_{t+1})] \le \mathcal{L}(\theta_t) - \eta_t\,\langle \nabla_{k(t)}\mathcal{L},\, \mathbb{E}[\hat{\boldsymbol{g}}_t]\rangle + \frac{\eta_t^2 L}{2}\,\mathbb{E}\big[\|\hat{\boldsymbol{g}}_t\|^2\big]. \tag{33}$$

From the bias bound in Lemma C.6: $\mathbb{E}[\hat{\boldsymbol{g}}_t] = \nabla_{k(t)}\mathcal{L} + b_t$, where $\|b_t\| = \mathcal{O}(\mu^2 L\, q)$. Therefore:

$$\begin{aligned}\langle \nabla_{k(t)}\mathcal{L},\, \mathbb{E}[\hat{\boldsymbol{g}}_t]\rangle &= \|\nabla_{k(t)}\mathcal{L}\|^2 + \langle \nabla_{k(t)}\mathcal{L},\, b_t\rangle \\ &\ge \|\nabla_{k(t)}\mathcal{L}\|^2 - \|\nabla_{k(t)}\mathcal{L}\|\,\|b_t\| \qquad \text{(Cauchy–Schwarz)},\end{aligned} \tag{34}$$

where $\|b_t\| \le C_b\,\mu^2 L\, q$ for an absolute constant $C_b > 0$.

Substituting into (33):

$$\mathbb{E}[\mathcal{L}(\theta_{t+1})] \le \mathcal{L}(\theta_t) - \eta_t\,\|\nabla_{k(t)}\mathcal{L}\|^2 + \eta_t\,C_b\,\mu^2 L\, q\,\|\nabla_{k(t)}\mathcal{L}\| + \frac{\eta_t^2 L}{2}\,\mathbb{E}\big[\|\hat{\boldsymbol{g}}_t\|^2\big]. \tag{35}$$

Step 4: Bounding $\mathbb{E}\big[\|\hat{\boldsymbol{g}}_t\|^2\big]$.

From the second moment bound in Lemma C.6:

$$\mathbb{E}\big[\|\hat{\boldsymbol{g}}_t\|^2\big] \le (q+3)\,\|\nabla_{k(t)}\mathcal{L}\|^2 + C_1\,\mu^2 L^2\, q^3 + \frac{C_2\,\sigma_\xi^2\, q}{2\mu^2}. \tag{36}$$

(The leading coefficient $(q+3)$ is the Gaussian fourth-moment coefficient $(q+2)$ for the pure signal term, plus an additional $+1$ that absorbs signal–bias cross terms from the $\boldsymbol{z}$-dependent Hessian remainder; see Lemma C.6.)

Under $\gamma = \alpha r$, we have $\|\nabla_{k(t)}\mathcal{L}\|^2 = \alpha^2\|\boldsymbol{g}_t\|^2$ by Part I, which does **not grow with** r. Therefore, all terms in the bound (36) are independent of r. Part II's descent estimate uses only the upper-bound half $\|\boldsymbol{g}_t\| = \mathcal{O}(1)$ of Assumption C.3, not the $\Theta(1)$ lower bound that Part I needs to make $\mathsf{SNR}_t = \Theta(1)$ a precise rank rate. Note that Assumption C.4 alone does *not* imply $\|\boldsymbol{g}_t\| = \mathcal{O}(1)$: the represented-adapter bound $\|(\gamma/r)\sum_k \boldsymbol{b}_k\boldsymbol{a}_k^\top\|_F = \mathcal{O}(1)$ controls a sum over all atoms with gauge freedom $(\boldsymbol{b}_k, \boldsymbol{a}_k) \mapsto (c\boldsymbol{b}_k, \boldsymbol{a}_k/c)$, while $\|\boldsymbol{g}_t\|^2 \le \|\boldsymbol{G}\|_2^2(\|\boldsymbol{a}_{k(t)}\|^2 + \|\boldsymbol{b}_{k(t)}\|^2)$ also depends

on the dense gradient $\boldsymbol{G}$ and on the active atom's individual factor norms. The two assumptions are therefore complementary and both used in Part II.

Step 5: Absorbing the cross-term via Young's inequality.

The cross-term in (35) is handled by Young's inequality ($ab \le \frac{a^2}{4} + b^2$):

$$\eta_t \, C_b \, \mu^2 L \, q \, \|\nabla_{k(t)}\mathcal{L}\| \le \frac{\eta_t}{4} \, \|\nabla_{k(t)}\mathcal{L}\|^2 + \eta_t \, C_b^2 \, \mu^4 L^2 \, q^2. \tag{37}$$

Substituting into (35):

$$\mathbb{E}[\mathcal{L}(\theta_{t+1})] \le \mathcal{L}(\theta_t) - \frac{3\eta_t}{4} \, \|\nabla_{k(t)}\mathcal{L}\|^2 + \eta_t \, C_b^2 \, \mu^4 L^2 \, q^2 + \frac{\eta_t^2 L}{2} \, \mathbb{E}\big[\|\hat{\boldsymbol{g}}_t\|^2\big] \, . \tag{38}$$

Step 6: Learning rate choice.

Choose a constant learning rate $\eta_t = \eta = \frac{c}{\sqrt{T}}$ for a constant $c > 0$. For sufficiently large T (specifically, $T \ge c^2 \cdot 4L^2(q+3)^2$), this satisfies:

$$\frac{\eta L}{2}(q+3) \le \frac{1}{4}, \qquad \text{i.e.,} \qquad \eta \le \frac{1}{2L(q+3)}. \tag{39}$$

This ensures that the $\frac{\eta^2 L}{2}(q+3)\|\nabla_{k(t)}\mathcal{L}\|^2$ term from (36) can be absorbed into the leading descent term.

Specifically, substituting (36) into (38):

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\theta_{t+1})] \le \mathcal{L}(\theta_t) &- \frac{3\eta}{4} \, \|\nabla_{k(t)}\mathcal{L}\|^2 + \frac{\eta^2 L}{2}(q+3) \, \|\nabla_{k(t)}\mathcal{L}\|^2 \\ &+ \eta \, C_b^2 \, \mu^4 L^2 \, q^2 + \frac{\eta^2 L}{2}\left(C_1 \, \mu^2 L^2 \, q^3 + \frac{C_2 \, \sigma_\xi^2 \, q}{2\mu^2}\right).\end{aligned} \tag{40}$$

Applying the learning rate constraint $\frac{\eta L}{2}(q+3) \le \frac{1}{4}$:

$$-\frac{3\eta}{4}\|\nabla_{k(t)}\mathcal{L}\|^2 + \frac{\eta}{4}\|\nabla_{k(t)}\mathcal{L}\|^2 = -\frac{\eta}{2}\|\nabla_{k(t)}\mathcal{L}\|^2. \tag{41}$$

This simplifies to:

$$\mathbb{E}[\mathcal{L}(\theta_{t+1})] \le \mathcal{L}(\theta_t) - \frac{\eta}{2} \, \|\nabla_{k(t)}\mathcal{L}\|^2 + \eta \, \Phi, \tag{42}$$

where Φ collects all terms independent of $\|\nabla_{k(t)}\mathcal{L}\|^2$:

$$\Phi \;=\; C_b^2 \, \mu^4 L^2 \, q^2 + \frac{\eta L}{2}\left(C_1 \, \mu^2 L^2 \, q^3 + \frac{C_2 \, \sigma_\xi^2 \, q}{2\mu^2}\right). \tag{43}$$

Step 7: Telescoping sum.

Rearranging (42) and summing from $t = 0$ to $T - 1$:

$$\frac{\eta}{2} \sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla_{k(t)}\mathcal{L}(\theta_t)\|^2\big] \le \mathcal{L}(\theta_0) - \mathcal{L}^* + T\eta \, \Phi. \tag{44}$$

Dividing both sides by $T\eta/2$:

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\big[\|\nabla_{k(t)}\mathcal{L}(\theta_t)\|^2\big] \le \frac{2(\mathcal{L}(\theta_0) - \mathcal{L}^*)}{T\eta} + 2\Phi. \tag{45}$$

Step 8: Final bound.

Substituting $\eta = c/\sqrt{T}$ into (45):

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla_{k(t)}\mathcal{L}(\theta_t)\|^2\big] \le \underbrace{\frac{2(\mathcal{L}(\theta_0)-\mathcal{L}^*)}{c\sqrt{T}}}_{\mathcal{O}(1/\sqrt{T})} + 2\Phi. \tag{46}$$

Expanding Φ from (43) with $\eta = c/\sqrt{T}$:

$$\begin{aligned} 2\Phi &= 2C_b^2\,\mu^4 L^2\,q^2 + \frac{cL}{\sqrt{T}}\left(C_1\,\mu^2 L^2\,q^3 + \frac{C_2\,\sigma_\xi^2\,q}{2\mu^2}\right) \\ &= \underbrace{\mathcal{O}(\mu^4 L^2\,q^2)}_{\text{smoothing bias (persistent)}} + \underbrace{\mathcal{O}\left(\frac{\mu^2 L^3\,q^3}{\sqrt{T}}\right)}_{\text{smoothing–variance cross-term}\,\to\,0} + \underbrace{\mathcal{O}\left(\frac{\sigma_\xi^2\,q}{T^{1/2}\mu^2}\right)}_{\text{residual noise}\,\to\,0}. \end{aligned} \tag{47}$$

For fixed μ, L, q, σ_ξ, the middle cross-term is $\mathcal{O}(1/\sqrt{T})$ and is therefore absorbed into the leading $\mathcal{O}(1/\sqrt{T})$ optimization term. The final convergence bound is:

$$\boxed{\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[\|\nabla_{k(t)}\mathcal{L}(\theta_t)\|^2\big] \le \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}\left(\mu^4 L^2\,q^2 + \frac{\sigma_\xi^2\,q}{T^{1/2}\mu^2}\right).} \tag{48}$$

Consequences.

1. The bound depends on $q = d_{\text{out}} + d_{\text{in}}$ but **not on the total rank** r, confirming that topology-aware scaling removes rank-induced signal degradation.
2. The persistent bias $\mathcal{O}(\mu^4 L^2 q^2)$ is controlled by the smoothing parameter μ. Choosing $\mu = \mathcal{O}(q^{-1/2})$ makes this $\mathcal{O}(L^2/q^0) = \mathcal{O}(L^2)$.
3. The residual noise term $\mathcal{O}(\sigma_\xi^2\,q/(T^{1/2}\mu^2))$ vanishes as $T \to \infty$.
4. As $T \to \infty$ and $\mu \to 0$ at an appropriate rate, the bound reduces to the standard non-convex ZO convergence rate $\mathcal{O}(1/\sqrt{T})$.

□

C.7 Proof of Corollary 4.5: Coverage Cost for Full-Adapter Stationarity

Proof. The atom factor-coordinate blocks are disjoint in the coordinates of $(\boldsymbol{B}, \boldsymbol{A})$, so full-adapter stationarity is measured by the aggregate quantity $\sum_{k=1}^{r}\|\nabla_k\mathcal{L}(\theta)\|^2$ already used in the statement. If $k(t)$ is sampled uniformly from $\{1,\dots,r\}$, taking expectation over the active atom gives

$$\mathbb{E}_{k(t)}\big[\|\nabla_{k(t)}\mathcal{L}(\theta)\|^2\big] = \frac{1}{r}\sum_{k=1}^{r}\|\nabla_k\mathcal{L}(\theta)\|^2.$$

For a deterministic cyclic schedule, the same identity holds after averaging over one full cycle. Thus full-adapter stationarity requires atom coverage, but this coverage factor is separate from the single-query ZO estimator dimension. □

Remark C.8 (Rank and coverage). Theorems 4.3 and 4.4 form a complete theoretical picture:

- Theorem 4.3 shows that under naive scaling, the ZO optimization quality *collapses* as rank increases.
- Theorem 4.4 shows that topology-aware scaling $\gamma = \alpha r$ *restores* the signal to a rank-invariant level, yielding a finite-difference measurement whose active FD-SNR no longer decays with rank.

- The convergence part of Theorem 4.4 places the restored signal inside a projected-gradient bound governed by the active dimension, the smoothing parameter μ, and the noise level σ_ξ.

Consequently, increasing rank no longer directly degrades the active finite-difference measurement once the atom is queried. Full-adapter stationarity still depends on the atom schedule and coverage over a cycle, which is why the main text states the gain as a per-query and per-cycle estimator-budget improvement rather than a worst-case global convergence dominance claim.

C.8 Proof of Corollary 4.6: Rank-m Blocks and Other Factorizations

The corollary uses α as the target coefficient for the active rank-m block. Under this block-level convention, the active block contains m atoms and the compensation rule becomes $\gamma = \alpha r/m$. The AR1-ZO specialization corresponds to $m = 1$, which recovers $\gamma = \alpha r$.

Derivation. **Step 1: Block-m forward pass.**

Consider the factored update partitioned into $\lceil r/m \rceil$ blocks, each containing m rank-1 atoms. The full update to the weight matrix is:

$$\Delta \boldsymbol{W} = \frac{\gamma}{r} \sum_{k=1}^{r} \boldsymbol{b}_k \boldsymbol{a}_k^\top. \tag{49}$$

At step t, let B_t be the active block of m atoms. Only the atoms in B_t are perturbed and updated. The effective contribution of the active block to the forward pass is:

$$\Delta \boldsymbol{W}_{\text{active}} = \frac{\gamma}{r} \sum_{k \in B_t} \boldsymbol{b}_k \boldsymbol{a}_k^\top. \tag{50}$$

Step 2: Explicit block gradient via Lemma 1.

By applying Lemma C.5 to each atom $k \in B_t$, the active-block gradient for the entire block is the concatenation:

$$\nabla_{B_t} \mathcal{L} = \frac{\gamma}{r} \bigoplus_{k \in B_t} \begin{pmatrix} \boldsymbol{G}\boldsymbol{a}_k \\ \boldsymbol{G}^\top \boldsymbol{b}_k \end{pmatrix}. \tag{51}$$

Its squared norm is:

$$\|\nabla_{B_t} \mathcal{L}\|^2 = \left(\frac{\gamma}{r}\right)^2 \sum_{k \in B_t} \left(\|\boldsymbol{G}\boldsymbol{a}_k\|^2 + \|\boldsymbol{G}^\top \boldsymbol{b}_k\|^2\right). \tag{52}$$

The summation term depends only on $\boldsymbol{G}$ and the atoms in B_t; all dependence on γ and r is isolated in the prefactor $(\gamma/r)^2$.

Step 3: Signal restoration condition.

The key observation is that each active atom contributes to the forward pass with coefficient γ/r. Under naive scaling ($\gamma = \alpha$), the per-atom coefficient is α/r, so the block of m atoms has total effective coefficient $m \cdot \alpha/r$. In a joint update of all r atoms simultaneously (the full-rank reference), the total effective coefficient is $r \cdot \alpha/r = \alpha$.

To match the full-rank reference signal, we require the block's total effective coefficient to equal α:

$$\frac{\gamma}{r} m = \alpha \quad \Longrightarrow \quad \gamma = \frac{\alpha r}{m}. \tag{53}$$

Step 4: Verification.

Substituting $\gamma = \alpha r/m$ into the explicit formula from Step 2:

$$\|\nabla_{B_t} \mathcal{L}\|^2 = \left(\frac{\alpha r/m}{r}\right)^2 \sum_{k \in B_t} \left(\|\boldsymbol{G}\boldsymbol{a}_k\|^2 + \|\boldsymbol{G}^\top \boldsymbol{b}_k\|^2\right) = \frac{\alpha^2}{m^2} \sum_{k \in B_t} \left(\|\boldsymbol{G}\boldsymbol{a}_k\|^2 + \|\boldsymbol{G}^\top \boldsymbol{b}_k\|^2\right), \tag{54}$$

which is independent of the ambient LoRA rank r for a fixed block size m. Equivalently, the total active block coefficient is $(\gamma/r)m = \alpha$, so the block-level finite-difference signal does not decay as the adapter rank grows. Setting $m = 1$ gives the rank-1 AR1-ZO rule $\gamma = \alpha r$. □

Remark C.9 (Scope of tensorized extensions). The same compensation logic applies to CP, Tucker, or other multilinear adapters once the joint-normalized parameterization and the sequential active block are specified. Adapter-specific proofs instantiate the corresponding complete block, tangent space, normalization, and active scale, but the topology-scaling principle is the same.

C.9 Proof of Theorem 4.7: Conditional Progress Advantage of Aligned Rank-1 Atoms

We first state the additional assumptions required beyond Assumptions C.1–C.3.

Assumption C.10 (Spectral Concentration)**.** The dense matrix gradient $\boldsymbol{G} = \nabla_{\boldsymbol{W}} \mathcal{L} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ has SVD $\boldsymbol{G} = \sum_i \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ with spectral concentration parameter $\rho = \sigma_1^2 / \|\boldsymbol{G}\|_F^2 \in (0, 1)$.

Assumption C.11 (Atom Alignment)**.** The active rank-1 atom $(\boldsymbol{b}, \boldsymbol{a})$ has alignment $\beta = \cos^2(\boldsymbol{b}, \boldsymbol{u}_1)\, \cos^2(\boldsymbol{a}, \boldsymbol{v}_1) > 0$ with the top singular pair $(\boldsymbol{u}_1, \boldsymbol{v}_1)$ of $\boldsymbol{G}$.

Theorem 4.7 (Restated). *Under Assumptions C.1, C.2, C.3, C.10, C.11, the one-step expected loss decrease admits the active-energy bound*

$$\mathbb{E}[-\Delta\mathcal{L}] \;\gtrsim\; \eta \Big(\frac{\gamma}{r}\Big)^2 \|\boldsymbol{g}_k\|^2 - \frac{\eta^2 L}{2}\, \mathbb{E}\Big[\|\hat{\nabla}_k \mathcal{L}\|^2\Big] - \eta\, \mathcal{O}(\mu^4 L^2\, q^2), \tag{55}$$

and the spectral lower bound

$$\|\boldsymbol{g}_k\|^2 \;\geq\; \rho\beta\, \|\boldsymbol{G}\|_F^2 \big(\|\boldsymbol{a}_k\|^2 + \|\boldsymbol{b}_k\|^2\big). \tag{56}$$

Assumption C.3 guarantees $\|\boldsymbol{a}_{k(t)}\|^2 + \|\boldsymbol{b}_{k(t)}\|^2 = \Theta(1)$ on the queried atom, so the active query carries a $\rho\beta$-scaled fraction of the dense gradient energy at active dimension q.

Proof. **Step 1: Active-atom gradient via SVD decomposition.**

By Lemma C.5, the active-atom gradient for atom k is:

$$\nabla_{\boldsymbol{b}} \mathcal{L} = \frac{\gamma}{r}\, \boldsymbol{G}\boldsymbol{a}_k, \qquad \nabla_{\boldsymbol{a}} \mathcal{L} = \frac{\gamma}{r}\, \boldsymbol{G}^\top \boldsymbol{b}_k. \tag{57}$$

Expanding $\boldsymbol{G}$ via its SVD $\boldsymbol{G} = \sum_i \sigma_i\, \boldsymbol{u}_i \boldsymbol{v}_i^\top$:

$$\boldsymbol{G}\boldsymbol{a}_k = \sum_i \sigma_i (\boldsymbol{v}_i^\top \boldsymbol{a}_k) \boldsymbol{u}_i, \qquad \boldsymbol{G}^\top \boldsymbol{b}_k = \sum_i \sigma_i (\boldsymbol{u}_i^\top \boldsymbol{b}_k) \boldsymbol{v}_i. \tag{58}$$

Step 2: Gradient energy decomposition.

The squared norms of the individual gradients are:

$$\|\boldsymbol{G}\boldsymbol{a}_k\|^2 = \sum_i \sigma_i^2 (\boldsymbol{v}_i^\top \boldsymbol{a}_k)^2, \qquad \|\boldsymbol{G}^\top \boldsymbol{b}_k\|^2 = \sum_i \sigma_i^2 (\boldsymbol{u}_i^\top \boldsymbol{b}_k)^2, \tag{59}$$

where we used the orthonormality of $\{\boldsymbol{u}_i\}$ and $\{\boldsymbol{v}_i\}$, respectively.

The total active-atom gradient energy is:

$$\|\nabla_k \mathcal{L}\|^2 = \Big(\frac{\gamma}{r}\Big)^2 \big(\|\boldsymbol{G}\boldsymbol{a}_k\|^2 + \|\boldsymbol{G}^\top \boldsymbol{b}_k\|^2\big). \tag{60}$$

Step 3: Lower bound via spectral concentration and alignment.

We lower-bound each term using Assumptions C.10 and C.11.

For the first term, retaining only the $i = 1$ contribution:

$$\|\boldsymbol{G}\boldsymbol{a}_k\|^2 = \sum_i \sigma_i^2 (\boldsymbol{v}_i^\top \boldsymbol{a}_k)^2 \geq \sigma_1^2 (\boldsymbol{v}_1^\top \boldsymbol{a}_k)^2 = \sigma_1^2 \cos^2(\boldsymbol{a}_k, \boldsymbol{v}_1) \|\boldsymbol{a}_k\|^2. \tag{61}$$

(The last equality uses $(\boldsymbol{v}_1^\top \boldsymbol{a}_k)^2 = \cos^2(\boldsymbol{a}_k, \boldsymbol{v}_1)\, \|\boldsymbol{a}_k\|^2$ for general (non-unit) $\boldsymbol{a}_k$. The factor norms $\|\boldsymbol{a}_k\|, \|\boldsymbol{b}_k\|$ are kept explicit; under Assumption C.3 they are $\Theta(1)$ on the queried atom, so the lower bound has the form $\Theta(1) \cdot \rho\beta \|\boldsymbol{G}\|_F^2$ in r.)

By Assumption C.10, $\sigma_1^2 \geq \rho \|\boldsymbol{G}\|_F^2$, so:

$$\|\boldsymbol{G}\boldsymbol{a}_k\|^2 \geq \rho \|\boldsymbol{G}\|_F^2 \cos^2(\boldsymbol{a}_k, \boldsymbol{v}_1)\|\boldsymbol{a}_k\|^2. \tag{62}$$

Similarly:

$$\|\boldsymbol{G}^\top \boldsymbol{b}_k\|^2 \geq \rho \|\boldsymbol{G}\|_F^2 \cos^2(\boldsymbol{b}_k, \boldsymbol{u}_1)\|\boldsymbol{b}_k\|^2. \tag{63}$$

Define $c_a = \cos^2(\boldsymbol{a}_k, \boldsymbol{v}_1)$, $c_b = \cos^2(\boldsymbol{b}_k, \boldsymbol{u}_1)$, and $\beta = c_a \cdot c_b$ (Assumption C.11). Combining:

$$\|\boldsymbol{G}\boldsymbol{a}_k\|^2 + \|\boldsymbol{G}^\top \boldsymbol{b}_k\|^2 \geq \rho \|\boldsymbol{G}\|_F^2 (c_a\|\boldsymbol{a}_k\|^2 + c_b\|\boldsymbol{b}_k\|^2). \tag{64}$$

Since $\beta = c_a c_b$ and $c_a, c_b \in [0, 1]$, we have $\beta \leq c_a$ and $\beta \leq c_b$. Hence

$$\|\boldsymbol{G}\boldsymbol{a}_k\|^2 + \|\boldsymbol{G}^\top \boldsymbol{b}_k\|^2 \geq \rho\beta \|\boldsymbol{G}\|_F^2 \left(\|\boldsymbol{a}_k\|^2 + \|\boldsymbol{b}_k\|^2\right). \tag{65}$$

Substituting into (60):

$$\|\nabla_k \mathcal{L}\|^2 \gtrsim \left(\frac{\gamma}{r}\right)^2 \rho\beta \|\boldsymbol{G}\|_F^2 \left(\|\boldsymbol{a}_k\|^2 + \|\boldsymbol{b}_k\|^2\right). \tag{66}$$

Step 4: Per-step descent via the Descent Lemma.

By the descent inequality established in the proof of Theorem 4.4 (eq. (33)), the expected loss decrease satisfies:

$$\mathbb{E}[-\Delta\mathcal{L}] = \mathbb{E}[\mathcal{L}(\theta_t) - \mathcal{L}(\theta_{t+1})] \geq \eta \langle \nabla_{k(t)}\mathcal{L}, \mathbb{E}[\hat{\boldsymbol{g}}_t]\rangle - \frac{\eta^2 L}{2} \mathbb{E}\left[\|\hat{\boldsymbol{g}}_t\|^2\right]. \tag{67}$$

From the bias bound in Lemma C.6:

$$\langle \nabla_{k(t)}\mathcal{L}, \mathbb{E}[\hat{\boldsymbol{g}}_t]\rangle = \|\nabla_{k(t)}\mathcal{L}\|^2 + \langle \nabla_{k(t)}\mathcal{L}, b_t\rangle \geq \|\nabla_{k(t)}\mathcal{L}\|^2 - \|\nabla_{k(t)}\mathcal{L}\| \, \mathcal{O}(\mu^2 L\, q). \tag{68}$$

Therefore:

$$\mathbb{E}[-\Delta\mathcal{L}] \geq \eta \|\nabla_{k(t)}\mathcal{L}\|^2 - \eta \, \mathcal{O}(\mu^2 L\, q) \|\nabla_{k(t)}\mathcal{L}\| - \frac{\eta^2 L}{2} \mathbb{E}\left[\|\hat{\boldsymbol{g}}_t\|^2\right]. \tag{69}$$

We handle the cross-term using Young's inequality ($ab \leq \frac{a^2}{4\epsilon} + \epsilon b^2$ with $\epsilon = 1$):

$$\eta C_b \mu^2 L q \|\nabla_{k(t)}\mathcal{L}\| \leq \frac{\eta}{4}\|\nabla_{k(t)}\mathcal{L}\|^2 + \eta C_b^2 \mu^4 L^2 q^2. \tag{70}$$

Absorbing the $\frac{\eta}{4}\|\nabla\|^2$ term into the leading descent:

$$\mathbb{E}[-\Delta\mathcal{L}] \geq \frac{3\eta}{4} \|\nabla_{k(t)}\mathcal{L}\|^2 - \eta \, C_b^2 \mu^4 L^2 \, q^2 - \frac{\eta^2 L}{2} \mathbb{E}\left[\|\hat{\boldsymbol{g}}_t\|^2\right]. \tag{71}$$

Substituting the lower bound (66) into the leading term:

$$\boxed{\mathbb{E}[-\Delta\mathcal{L}] \gtrsim \eta\left(\frac{\gamma}{r}\right)^2 \|\boldsymbol{g}_k\|^2 - \frac{\eta^2 L}{2} \mathbb{E}\left[\|\hat{\boldsymbol{g}}_t\|^2\right] - \eta \, \mathcal{O}(\mu^4 L^2 \, q^2).} \tag{72}$$

The lower bound on $\|\boldsymbol{g}_k\|^2$ from Step 3 gives the sufficient spectral-alignment version stated in the main text. □

Remark C.12 (Comparison with generic random blocks). For a *generic* coordinate block B of the same dimension $q = d_{\text{out}} + d_{\text{in}}$, a standard random-block averaging calculation gives:

$$\mathbb{E}_B\left[\|\nabla_B \mathcal{L}\|^2\right] = \frac{q}{d_{\text{out}} d_{\text{in}}} \|\boldsymbol{G}\|_F^2. \tag{73}$$

The advantage ratio of the aligned rank-1 atom over a generic block is therefore:

$$\frac{\rho\beta \|\boldsymbol{G}\|_F^2 \left(\|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2\right)}{\frac{q}{d_{\text{out}} d_{\text{in}}} \|\boldsymbol{G}\|_F^2} = \rho\beta \, \frac{d_{\text{out}} d_{\text{in}}}{d_{\text{out}} + d_{\text{in}}} \left(\|\boldsymbol{a}\|^2 + \|\boldsymbol{b}\|^2\right). \tag{74}$$

Under Assumption C.3 (active factor norms $\|\boldsymbol{b}_{k(t)}\|, \|\boldsymbol{a}_{k(t)}\| = \Theta(1)$) and $d_{\text{out}} = d_{\text{in}} = d$, the advantage ratio simplifies to

$$\frac{\rho\beta \|\boldsymbol{G}\|_F^2 (\|\boldsymbol{a}_{k(t)}\|^2 + \|\boldsymbol{b}_{k(t)}\|^2)}{(q/(d_{\text{out}} d_{\text{in}})) \|\boldsymbol{G}\|_F^2} = \rho\beta \, \frac{d_{\text{out}} d_{\text{in}}}{q} \left(\|\boldsymbol{a}_{k(t)}\|^2 + \|\boldsymbol{b}_{k(t)}\|^2\right) = \Theta(\rho\beta \cdot d).$$

At $d = 4096$, $\rho = 0.1$, $\beta = 0.05$ with bounded factor norms, the sufficient lower-bound ratio is $\Theta(0.1 \cdot 0.05 \cdot 4096) \approx \Theta(20)$.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Introduction state the paper's scope as a topology-aware rank-1 ZO query method for LoRA fine-tuning, with theory on atom minimality, active query dimension, directional collapse/restoration, and coverage cost. The downstream claims are tied to the OPT/Qwen3 matched-budget experiments, with limitations stated in the conclusion.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion includes a limitations and future work paragraph noting that the main stationarity result is active-atom stationarity, that full-adapter stationarity still pays an atom-coverage cost, and that experiments are limited to OPT/Qwen3 families and selected tasks. The theory section and appendix also state the active-gradient and represented-adapter-scale assumptions used by the results.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The main text states the formal theorem, lemma, proposition, and corollary results with explicit references to the needed assumptions. Appendix C provides the assumptions, auxiliary lemmas, and full proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.

- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5.1 gives the models, tasks, baselines, rank/scaling choices, and matched two-forward-pass evaluation contract. Appendix B provides task splits, baseline tuning, diagnostic protocols, ablation settings, and profiling details needed to reproduce the reported comparisons.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The submission includes an anonymized code package with the implementation, experiment scripts, and instructions needed to reproduce the main results. The paper uses public models, datasets, and baseline implementations where available, with data access and baseline settings described in the appendix and supplemental material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5.1 summarizes the experimental setup, and Appendix B gives the task splits, baseline tuning policy, rank/scaling choices, smoothing radius, batch size, seeds, profiling protocol, and diagnostic settings.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports variability information through error bars or bands in the relevant tables and plots, together with matched-budget downstream scores, loss trajectories, and mechanism diagnostics. The experimental section and appendix describe the corresponding evaluation settings and aggregation protocols.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix B reports the GPU types used for the OPT and Qwen3 experiments, and the efficiency study reports step time and peak-memory profiling under the matched protocol. These details provide the compute and memory envelope needed to reproduce the reported runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work studies an optimization method for memory-efficient fine-tuning using public benchmarks and pretrained models, and does not involve human-subject data collection, deception, privacy-invasive data gathering, or deployment claims. The submission is anonymized for double-blind review.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The work may have positive impact by reducing memory barriers for LLM fine-tuning, making experimentation more accessible and energy-efficient. As with other efficient fine-tuning methods, it could also lower the cost of adapting models for harmful or low-quality applications; the paper does not release a new high-risk model or dataset.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not introduce or release a new pretrained language model, image generator, scraped dataset, or other asset with direct high-risk misuse potential. It proposes an optimization method evaluated on existing public models and benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the original works for the pretrained models, datasets, and baseline methods used in the experiments, and the authors respect the corresponding public terms of use. The submitted/released materials are licensed under the Creative Commons Attribution 4.0 International license.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The submission introduces an anonymized code package as a new asset, including the AR1-ZO implementation, experiment scripts, reproduction instructions, and license information. No new dataset, pretrained model, or benchmark is introduced.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, new data collection from people, or experiments with human subjects. All experiments use existing public benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human-subject or crowdsourcing study is conducted, so IRB approval or equivalent review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are the experimental objects being fine-tuned, but no LLM is used as an important, original, or non-standard component of the method development, data generation, evaluation, or scientific claims. Any use of language tools for writing or formatting would not affect the core methodology or results.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.