Preserve then Quantize: Dominant-Subspace Guided Low-Rank Reconstruction

Yoonjun Cho [1 *] **Dongjae Jeon** [1 *] **Soeun Kim** [1] **Albert No** [1]

Abstract

Post-training quantization (PTQ) enables efficient deployment of LLMs by converting weights to low-bit formats, but often degrades accuracy. Quantization error reconstruction (QER) mitigates this by adding a low-rank correction term. However, existing QER methods typically quantize weights before identifying low-rank structure, discarding information they later attempt to recover. We propose *Structured Residual Reconstruction (SRR)*, a simple yet effective reformulation of QER that first preserves dominant spectral directions and quantizes only the residual tail. The final approximation combines the preserved low-rank structure with a quantized residual, yielding improved fidelity under the same rank constraint. SRR generalizes to activation-aware settings by selecting dominant components based on contributions in both the original and activation-weighted spaces. We also apply SRR in QPEFT by freezing the preserved subspace and updating only the residual component during fine-tuning, which stabilizes training and leads to better adaptation. Across both PTQ and QPEFT, SRR consistently improves performance under fixed rank constraints, providing an effective framework for quantization-aware compression.

1. Introduction

Post-Training Quantization (PTQ) reduces the memory and computational costs of LLMs by converting weights to low-bit formats (Nagel et al., 2021; Frantar et al., 2023), but this often leads to a significant drop in accuracy (Yao et al., 2022; Lin et al., 2024; Tseng et al., 2024). Quantization Error Reconstruction (QER) addresses this by adding a low-rank correction term, resulting in an approximation of the form $\mathbf{W} \approx \mathbf{Q} + \mathbf{LR}$ that aims to recover information lost during quantization (Yao et al., 2024; Zhang et al., 2024a; Liu et al., 2024; Zhang et al., 2025).

While most QER methods assume that quantization error can be effectively captured by a low-rank matrix (Zhang et al., 2024a; 2025), this assumption does not always hold in practice. The error spectrum can vary across layers, and truncated SVD can capture only a portion of its variance. In addition, QER methods typically apply quantization before identifying low-rank structure, which might result in the loss of information that could otherwise be retained.

To address this limitation, we propose *Structured Residual Reconstruction (SRR)*, a simple yet effective reformulation of QER that captures dominant components of weight *before* quantization. Instead of quantizing the full weight matrix, SRR separates dominant and tail components, quantizes only the latter, and reconstructs the approximation via truncated SVD of their combination. This design prevents early information loss and empirically yields more accurate low-rank approximations. In activation-aware settings, we generalize SRR by selecting directions to preserve based on their contributions in both the original space $\mathbf{W}$ and the scaled space $\mathbf{SW}$, allowing adaptive preservation of informative subspaces.

We further seamlessly integrate SRR into the QPEFT setting by partially freezing the dominant directions during fine-tuning and updating only the residual subspace. This strategy stabilizes learning while preserving the structural integrity of the approximation, leading to improved performance. Through extensive experiments under both PTQ and QPEFT settings, we demonstrate that SRR outperforms existing methods in both quantization and fine-tuning tasks.

To summarize, our contributions are as follows:

- We propose *Structured Residual Reconstruction (SRR)*, a quantization scheme that preserves dominant directions before quantization to improve approximation.
- We generalize SRR to activation-aware settings and integrate it into QPEFT by freezing preserved subspaces during fine-tuning.
- SRR consistently outperforms existing QER methods under fixed-rank constraints in both PTQ and QPEFT.

[*] Equal contribution [1] Yonsei University. Correspondence to: Albert No <albertno@yonsei.ac.kr>.

TTODLer-FM Workshop at ICML 2025. Copyright 2025 by the author(s).

2. Preliminaries

Quantization Error Reconstruction. Post-Training Quantization (PTQ) compresses model weights into low-bit representations (Banner et al., 2019; Nagel et al., 2020), but often leads to accuracy degradation due to quantization-induced information loss. Quantization Error Reconstruction (QER) mitigates this issue by introducing a low-rank correction term that approximates the discrepancy between the original weights and their quantized counterparts (Yao et al., 2024; Zhang et al., 2024a; Liu et al., 2024; Zhang et al., 2025). This approach restores lost expressiveness with minimal overhead, significantly improving the performance of quantized models.

Formally, consider a linear transformation $\mathbf{y} = \mathbf{x}\mathbf{W}$, where $\mathbf{W} \in \mathbb{R}^{m\times n}$ is a full-precision weight matrix and $\mathbf{x} \in \mathbb{R}^{m}$ is the input. Its quantized counterpart is denoted by $\mathbf{Q}$, representing the low-bit approximation of $\mathbf{W}$. QER restores the output accuracy by adding a rank-r correction term $\mathbf{LR}$:

$$\mathbf{y}_q = \mathbf{x}(\mathbf{Q} + \mathbf{LR}) \approx \mathbf{x}\mathbf{W},$$

where $\mathbf{L} \in \mathbb{R}^{m\times r}$, $\mathbf{R} \in \mathbb{R}^{r\times n}$, and $r \ll \min(m, n)$.

The correction term is typically computed by applying rank-r truncated singular value decomposition (SVD) to the residual:

$$\mathbf{W} - \mathbf{Q} \approx \mathbf{LR}, \quad \text{with } \mathbf{LR} = \mathrm{SVD}_r(\mathbf{W} - \mathbf{Q}).$$

ZeroQuant-V2 (Yao et al., 2024) constructs the low-rank correction term using truncated SVD of the residual $\mathbf{W}-\mathbf{Q}$, but does not account for input distributions, which may limit its robustness under varying activation statistics.

To better align reconstruction with input-dependent behavior, LQER (Zhang et al., 2024a), QERA-approx (Zhang et al., 2025) introduces a heuristic scaling matrix derived from calibration data. Building on this insight, QERA-exact (Zhang et al., 2025), EoRA (Liu et al., 2024), and CALDERA (Saha et al., 2024) derive exact closed-form solutions to the same problem: minimizing reconstruction error in the layer output $\mathbf{x}\mathbf{W}$. These methods compute a data-driven scaling matrix $\mathbf{S}$ that reweights the reconstruction objective to emphasize directions with higher output sensitivity.

Quantized Parameter-Efficient Fine-Tuning. Parameter-Efficient Fine-Tuning (PEFT) methods such as LoRA (Hu et al., 2022) adapt large pre-trained models to downstream tasks by inserting trainable low-rank adapters into frozen weights. These adapters are typically initialized to produce zero output and updated during fine-tuning.

Quantized Parameter-Efficient Fine-Tuning (QPEFT) extends this paradigm to quantized models, where adapter tuning must compensate for quantization-induced distortions. QLoRA (Dettmers et al., 2023) combines 4-bit quantization with LoRA, but the mismatch between dequantized and original weights makes zero-initialization less effective.

To mitigate this issue, LoftQ (Li et al., 2023) and LQ-LoRA (Guo et al., 2024) refine adapter initialization through iterative updates that alternate between quantizing weights and computing low-rank SVDs, aiming to approximate the original weights as $\mathbf{W} \approx \mathbf{Q} + \mathbf{LR}$; increasing the number of iterations generally leads to lower reconstruction error. LoftQ applies SVD after quantization without considering activation statistics, while LQ-LoRA performs scaled SVD beforehand using a Fisher-weighted objective to emphasize sensitive directions.

In contrast to iterative approaches, QERA (Zhang et al., 2025) applies an analytical QER formulation to QPEFT under the assumption that embedding dimensions are uncorrelated. Originally developed for PTQ, this method initializes adapters in QPEFT using a closed-form low-rank correction, without requiring iterative updates.

A Unified View of QER via Scaled Space. Recent QER methods can be unified under a common framework based on low-rank approximation in a *scaled space*, where the residual is pre-multiplied by a sensitivity-aware matrix $\mathbf{S} \in \mathbb{R}^{m\times m}$. Instead of directly approximating the quantization error $\mathbf{W} - \mathbf{Q}$, these methods perform rank-r truncated SVD on the scaled residual:

$$\mathbf{LR} = \mathbf{S}^{-1}\,\mathrm{SVD}_r\left(\mathbf{S}(\mathbf{W} - \mathbf{Q})\right).$$

The correction is then constructed by applying truncated SVD to the scaled residual $\mathbf{S}(\mathbf{W} - \mathbf{Q}) = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$, and projecting it back to the original space by multiplying $\mathbf{S}^{-1}$:

$$\mathbf{L} := \mathbf{S}^{-1}\mathbf{U}_{:,:r}, \quad \mathbf{R} := \boldsymbol{\Sigma}_{:r,:r}\mathbf{V}^\top_{:,:r}.$$

Each QER variant corresponds to a different choice of the scaling matrix $\mathbf{S}$. Among existing methods, ZeroQuant-V2 sets $\mathbf{S} = \mathbf{I}$ and minimizes unweighted reconstruction error. In contrast, methods such as LQER (Zhang et al., 2024a), QERA (Zhang et al., 2025), EoRA (Liu et al., 2024), and CALDERA (Saha et al., 2024) use $\mathbf{S} \neq \mathbf{I}$, incorporating input-dependent scaling.

This framework highlights a common structure across QER methods, where $\mathbf{S}$ defines the geometry of the reconstruction objective. The resulting error in scaled space is given by

$$\|\mathbf{S}(\mathbf{W} - \mathbf{Q} - \mathbf{LR})\|_F\,,$$

which measures the residual after projection onto the top-r directions in the sensitivity-weighted domain. This perspective motivates our method, which promotes alignment between quantization error and dominant directions under the chosen scaling.

3. Method

3.1. Motivation

The standard approach to Quantization Error Reconstruction (QER) (Zhang et al., 2025; Liu et al., 2024) begins by quantizing the full weight matrix $\mathbf{W}$, computing the quantization error $\mathrm{E}_q(\mathbf{W}) := \mathbf{W} - \mathcal{Q}(\mathbf{W})$, and then approximating this residual using low-rank terms $\mathbf{LR}$, typically obtained via truncated SVD with fixed rank r.

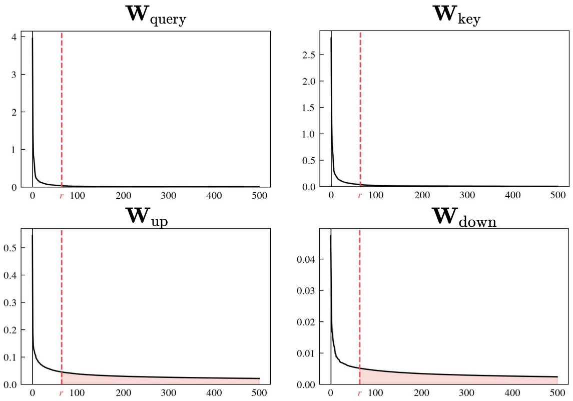

Figure 1. Singular value spectra of the scaled quantization error $\mathrm{E}_q(\mathbf{SW})$ for various weight matrices in the first layer of LLaMA-2 7B. The top-r components are retained by the low-rank correction; the pink region shows unrecovered residuals. Query and Key errors exhibit low-rank structure, while Up and Down projections show flatter spectra with limited recoverability. This pattern holds consistently across layers.

This pipeline implicitly assumes that the quantization error $\mathrm{E}_q(\mathbf{W})$, or its activation-scaled variant $\mathbf{S}\,\mathrm{E}_q(\mathbf{W})$, exhibits sufficient low-rank structure to be effectively captured via truncated SVD. However, we observe that this assumption often fails in practice. As shown in Figure 1, the spectral characteristics of $\mathrm{E}_q(\mathbf{W})$ vary widely across layers and matrix types. In particular, Up and Down projections often exhibit flat singular value spectra, indicating a lack of low-rank structure in the original matrix $\mathbf{W}$. In such cases, quantization may further degrade structure, making post-hoc low-rank recovery even less effective. Consequently, applying SVD to $\mathrm{E}_q(\mathbf{W})$ tends to capture only a small fraction of the error, especially when the spectrum is dispersed.

This reveals a limitation of existing QER approaches: by applying quantization to the entire matrix $\mathbf{W}$ before identifying low-rank structure, they risk discarding compressible structure that could otherwise be more faithfully recovered.

In this work, we propose a new decomposition framework that prioritizes the extraction of low-rank structure *directly from* $\mathbf{W}$, prior to any quantization. Rather than treating the low-rank term as a residual patch applied after quantization, we formulate a joint approximation of the form $\mathbf{W} \approx \mathbf{Q} + \mathbf{LR}$, minimizing the activation-weighted reconstruction error $\|\mathbf{S}(\mathbf{W} - \mathbf{Q} - \mathbf{LR})\|_F^2$. Our results demonstrate that allowing the low-rank component account for the dominant structure of $\mathbf{W}$ yields significantly better approximation quality, particularly in layers where quantization alone disrupts rank structure most severely.

3.2. A Simplified Case ($\mathbf{S} = \mathbf{I}$): Prioritizing Low-Rank Structure Before Quantization

To illustrate our key idea, we begin with a simplified setting where activation statistics are not considered (i.e., the scaling matrix is identity, $\mathbf{S} = \mathbf{I}$). In this case, our goal is to decompose the weight matrix $\mathbf{W}$ into two parts: 1) a quantized part $\mathbf{Q}$, which ideally has small norm to minimize quantization noise, and 2) a low-rank component $\mathbf{LR}$ that captures the dominant structure of $\mathbf{W}$.

Rather than quantizing the entire matrix as in conventional QER pipelines, we explicitly aim to preserve the high-energy directions in $\mathbf{W}$ through a low-rank representation and reserve quantization only for the low-energy residual.

Let the singular value decomposition (SVD) of $\mathbf{W}$ be:

$$\mathbf{W} = \mathbf{U}_h \mathbf{\Sigma}_h \mathbf{V}_h^\top + \mathbf{U}_\ell \mathbf{\Sigma}_\ell \mathbf{V}_\ell^\top,$$

where the first term corresponds to the top-r singular directions, and the second term represents the low-energy tail. We apply quantization only to the tail component:

$$\mathbf{Q} := \mathcal{Q}\left(\mathbf{U}_\ell \mathbf{\Sigma}_\ell \mathbf{V}_\ell^\top\right),$$

ensuring that the quantized component $\mathbf{Q}$ has relatively small norm and reduced quantization noise.

Next, we compute the structured residual $\mathbf{W} - \mathbf{Q}$ which retains both the unquantized top-r directions and the quantization error from the tail. To obtain the final low-rank approximation, we apply a second truncated SVD:

$$\begin{aligned}\mathbf{LR} &:= \mathrm{SVD}_r\left(\mathbf{W} - \mathbf{Q}\right) \\ &= \mathrm{SVD}_r\left(\mathbf{U}_h \mathbf{\Sigma}_h \mathbf{V}_h^\top + \mathrm{E}_q\left(\mathbf{U}_\ell \mathbf{\Sigma}_\ell \mathbf{V}_\ell^\top\right)\right).\end{aligned}$$

The final decomposition is therefore $\mathbf{W} \approx \mathbf{Q} + \mathbf{LR}$ which we refer to as *Structured Residual Reconstruction (SRR)*. As shown in Figure 2, SRR consistently outperforms standard QER across a range of model scales and layers.

Crucially, SRR is a simple reversal of the conventional QER strategy. While QER begins by quantizing the entire weight matrix $\mathbf{W}$ and then uses low-rank approximation to recover from the resulting error, SRR inverts this process: it first extracts the dominant low-rank structure and then applies quantization only to the residual. Though seemingly minor, this reversal offers key advantages by allowing the low-rank term to capture the structured directions of W directly, rather than forcing it to compensate for unstructured or high-magnitude quantization noise.

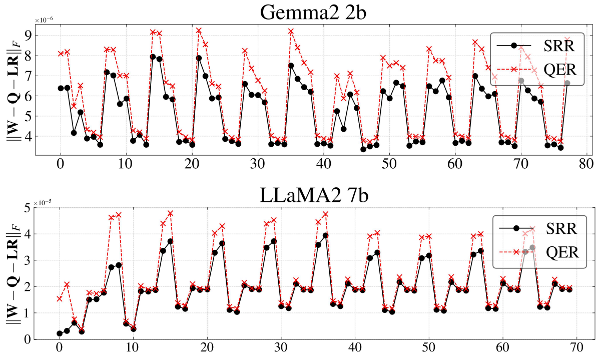

Figure 2. Layer-wise decomposition error (clipped for visibility) using SRR and QER under $\mathbf{S} = \mathbf{I}$. SRR consistently achieves lower error than QER in both Gemma-2 2B (top) and LLaMA-2 7B (bottom). Full results with layer names are in Figure 4.

Importantly, although SRR is introduced in the simplified setting where $\mathbf{S} = \mathbf{I}$, its core intuition readily generalizes. As we show in the next section, the same principle of prioritizing low-rank structure before quantization extends naturally to the general case where activation statistics are incorporated through a non-identity scaling matrix $\mathbf{S}$.

3.3. Extending SRR to the General Case $\mathbf{S} \neq \mathbf{I}$: Incorporating Activation Statistics

In the general case where activation statistics are incorporated through a non-identity scaling matrix $\mathbf{S}$, the low-rank term $\mathbf{LR}$ in SRR is defined as:

$$\begin{aligned}\mathbf{LR} &:= \mathbf{S}^{-1}\,\mathrm{SVD}_r\left(\mathbf{S}(\mathbf{W}-\mathbf{Q})\right)\\ &= \mathbf{S}^{-1}\,\mathrm{SVD}_r\left(\mathbf{U}_h\mathbf{\Sigma}_h\mathbf{V}_h^\top + \mathbf{S}\,\mathrm{E}_q\left(\mathbf{S}^{-1}\mathbf{U}_\ell\mathbf{\Sigma}_\ell\mathbf{V}_\ell^\top\right)\right).\end{aligned}$$

Interestingly, we observe that this naive extension of SRR does not consistently outperform QER. As shown in Figure 3, applying SRR in the scaled space can even lead to larger reconstruction errors than QER in certain layers.

We argue that this unexpected behavior arises from a subtle mismatch between the scaled space $\mathbf{SW}$ and the original parameter space $\mathbf{W}$. Specifically, the quantization input in this case is not simply the low-energy tail of $\mathbf{SW}$, but a rescaled version involving $\mathbf{S}^{-1}$ (see the E_q term above). This implies that directions which appear low in energy in the scaled space $\mathbf{SW}$ may contribute more significantly in the original space $\mathbf{W}$ after inverse mapping, ultimately leading to suboptimal decomposition.

To address this, we propose an adaptive strategy that selects low-rank components based on their significance in both the scaled space $\mathbf{SW}$ and the original space $\mathbf{W}$. This enables the retention of directions whose importance is preserved under inverse scaling, thereby mitigating the amplification of quantization errors in the original parameter space.

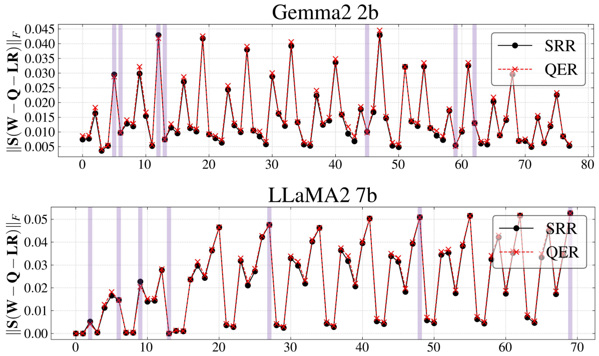

Figure 3. Layer-wise decomposition error (clipped) of SRR and QER under $\mathbf{S} \neq \mathbf{I}$ (QERA-exact). Naive application of SRR does not always outperform QER in either Gemma-2 2B (top) or LLaMA-2 7B (bottom). Layers where SRR performs worse are shown in purple. Full results with layer names are in Figure 5.

We begin by expressing the SVD of $\mathbf{SW}$ as:

$$\mathbf{SW} = \sum_{i=1}^{n} \sigma_i u_i v_i^\top,$$

where σ_i are singular values and u_i, v_i are left and right singular vectors. When these directions are mapped back to the original space, each contributes:

$$\mathbf{W}_i := \sigma_i \mathbf{S}^{-1} u_i v_i^\top,$$

with Frobenius norm $\|\mathbf{W}_i\|_F := \sigma_i \|\mathbf{S}^{-1}u_i\|_2$. This quantity reflects how strongly each direction in $\mathbf{SW}$ contributes to $\mathbf{W}$. Accordingly, we define a contribution score:

$$\mathrm{score}_i := \sigma_i \cdot \|\mathbf{S}^{-1}u_i\|_2,$$

which captures the relative importance of the i-th direction in the original space.

Using this quantity, we retain directions that are considered spectrally important in both spaces. Let $[r] := \{1, 2, \ldots, r\}$ denote the indices of the top-r directions ranked by singular value in the scaled space $\mathbf{SW}$. Separately, let Top-$r(\mathrm{score}_i)$ denote the indices of the r directions with the highest contribution scores. We define the preserved index set as

$$\mathcal{H} := [r] \cap \text{Top-}r(\mathrm{score}_i),$$

and the tail set as $\mathcal{L} := [n] \setminus \mathcal{H}$.

This leads to the following decomposition of $\mathbf{W}$:

$$\mathbf{W} = \underbrace{\sum_{i\in\mathcal{H}} \sigma_i \mathbf{S}^{-1} u_i v_i^\top}_{\text{preserved}} + \underbrace{\sum_{i\in\mathcal{L}} \sigma_i \mathbf{S}^{-1} u_i v_i^\top}_{\text{tail}},$$

or, in matrix form:

$$\mathbf{W} = \mathbf{S}^{-1}\mathbf{U}_{\mathcal{H}}\mathbf{\Sigma}_{\mathcal{H}}\mathbf{V}_{\mathcal{H}}^\top + \mathbf{S}^{-1}\mathbf{U}_{\mathcal{L}}\mathbf{\Sigma}_{\mathcal{L}}\mathbf{V}_{\mathcal{L}}^\top.$$

Note that this formulation reduces to SRR in which the top r directions are selected, when the scaling matrix is the identity ($\mathbf{S} = \mathbf{I}$). Importantly, it incurs negligible additional cost, as all quantities are derived from the SVD of $\mathbf{SW}$.

Building on this formulation, we apply the decomposition strategy to Post-Training Quantization (PTQ) by preserving the dominant directions $\mathcal{H}$ and quantizing the tail $\mathcal{L}$:

$$\mathbf{Q} := \mathcal{Q}(\mathbf{S}^{-1}\mathbf{U}_{\mathcal{L}}\mathbf{\Sigma}_{\mathcal{L}}\mathbf{V}_{\mathcal{L}}^{\top}), \quad \mathbf{LR} := \mathbf{S}^{-1}\,\mathrm{SVD}_r(\mathbf{S}(\mathbf{W}-\mathbf{Q})).$$

The final approximation is given by

$$\widehat{\mathbf{W}}_{\mathrm{PTQ}} := \mathbf{Q} + \mathbf{LR}.$$

This embodies the core idea of SRR: it preserves the essential structure of $\mathbf{W}$ through low-rank modeling, while also resolving the mismatch between $\mathbf{SW}$ and $\mathbf{W}$ by explicitly selecting directions that remain important after inverse scaling. As a result, it yields more accurate approximations under arbitrary scaling matrices $\mathbf{S}$ and across layers.

3.4. Application to QPEFT: Freezing Dominant Directions for Efficient Fine-Tuning

The decomposition $\widehat{\mathbf{W}} := \mathbf{Q} + \mathbf{LR}$ can be directly applied to Quantized Parameter-Efficient Fine-Tuning (QPEFT). In this setup, the quantized weights $\mathbf{Q}$ are kept fixed, and only the low-rank term $\mathbf{LR}$ is updated during fine-tuning. This aligns naturally with the LoRA framework, where $\mathbf{LR}$ serves as a trainable residual added to a frozen weight, enabling efficient adaptation with minimal cost.

In our approach, the low-rank term $\mathbf{LR}$ is initialized to approximate the dominant subspace of the weight matrix. While this decomposition proves effective for PTQ, directly fine-tuning the entire $\mathbf{LR}$ can degrade performance. Empirically, we observe that unconstrained updates often distort critical directions aligned with the top singular directions.

To address this, we propose a partial freezing strategy. We freeze the preserved top k directions, which correspond to the dominant subspace $\mathbf{U}_{\mathcal{H}}\mathbf{\Sigma}_{\mathcal{H}}\mathbf{V}_{\mathcal{H}}^{\top}$ within the low-rank term ($k \leq r$), and fine-tune only the remaining residual subspace. This approach preserves the principal structure encoded during quantization while allowing sufficient flexibility for downstream task adaptation.

Let k denote the number of frozen (preserved) directions. We decompose the rank-r low-rank component as follows:

$$\mathbf{L}^{\mathrm{frozen}} := \mathbf{S}^{-1}\mathbf{U}_{:,:k}\mathbf{\Sigma}_{:k,:k}, \quad \mathbf{R}^{\mathrm{frozen}} := \mathbf{V}_{:,:k}^{\top}$$
$$\mathbf{L}^{\mathrm{tune}} := \mathbf{S}^{-1}\mathbf{U}_{:,k:r}\mathbf{\Sigma}_{k:r,k:r}, \quad \mathbf{R}^{\mathrm{tune}} := \mathbf{V}_{:,k:r}^{\top}$$

Only $\mathbf{L}^{\mathrm{tune}}$ and $\mathbf{R}^{\mathrm{tune}}$ are trainable; the leading directions $\mathbf{L}^{\mathrm{frozen}}$ and $\mathbf{R}^{\mathrm{frozen}}$ are kept fixed throughout fine-tuning to maintain the structural integrity of the original model.

To ensure both stability and adaptability, we choose $k \leq r/2$, which reserves sufficient capacity in the residual space for learning. Setting k too large overly restricts adaptation, whereas setting $k = 0$ may disregard the structure preserved during quantization. This hybrid strategy bridges quantization-aware initialization with parameter-efficient adaptation, offering a robust solution for QPEFT.

4. Experiments

In this section, we evaluate the effectiveness of SRR across a broad spectrum of models, bitwidths, and baselines, under both PTQ and QPEFT. Complete experimental details are provided in Appendix A, with PTQ settings detailed in Appendix A.1 and QPEFT outlined in Appendix A.2.

4.1. Experiments on PTQ

Table 1 presents the perplexity of quantized LLMs on WikiText2 under the 3-bit PTQ setting using MXINT (Darvish Rouhani et al., 2023), evaluated at two low-rank configurations: $r = 8$ and $r = 64$. Our SRR consistently improves perplexity when applied on top of existing QER methods, confirming its effectiveness across model scales. Additional PTQ results and further analyses are provided in Appendix B.

4.2. Experiments on QPEFT.

Table 2 presents results on the eight GLUE (Wang et al., 2019) tasks under 4-, 3-, and 2-bit quantization using the MXINT quantizer setting. We apply SRR within the QPEFT framework by freezing the dominant subspace and updating only the residual. This structured adaptation consistently outperforms existing methods across all bitwidths, with gains especially pronounced at lower bitwidths such as 2-bit, where quantization artifacts severely degrade model capacity. By preserving dominant directions and fine-tuning only the residual subspace, our method maintains stability while enabling effective adaptation. We also outperform iterative methods such as LoftQ (Li et al., 2023) and LQ-LoRA (Guo et al., 2024), despite using a single-pass procedure.

This performance advantage also extends to reasoning tasks. As shown in Table 3, our method achieves the highest accuracy on GSM8K (Cobbe et al., 2021) across all quantization levels, demonstrating robustness in multi-step reasoning under strong compression.

Importance of Freezing dominant directions. To further validate the importance of freezing dominant directions we conduct an ablation study comparing two variants: 1) fine-tunes the entire low-rank term $\mathbf{LR}$, and 2) freezes the leading directions while fine-tuning only the residual subspace. Since our method explicitly stores important structural in-

Table 1. Perplexity (↓) on WikiText2 under 3-bit PTQ with MXINT. We apply SRR to various baselines across four models and two LR ranks (r=8, 64). Lower perplexity scores are highlighted in **bold**. All evaluations are conducted using `lm-eval`.

		Method	**TinyLlama 1.1B**		**Gemma-2 2B**		**LLaMA-2 7B**		**LLaMA-3.1 8B**	
			$r = 8$	$r = 64$	$r = 8$	$r = 64$	$r = 8$	$r = 64$	$r = 8$	$r = 64$
		BF16	13.98		13.08		8.71		7.55	
Quantization Bits	**3.25**	*w-only*	32.82		41.13		13.33		18.96	
		ZeroQuant-V2 (Yao et al., 2024)	**30.09**	25.90	37.74	33.09	**12.96**	**12.99**	20.79	19.28
		w/ SRR	33.89	**25.18**	**31.93**	**31.08**	13.05	13.30	**18.44**	**18.44**
		LQER (Zhang et al., 2024a)	24.06	20.63	29.11	21.37	15.59	15.14	13.44	11.90
		w/ SRR	**23.64**	**19.86**	**28.70**	**21.02**	**12.49**	**11.05**	**13.14**	**11.76**
		QERA-approx (Zhang et al., 2025)	23.65	20.52	27.65	21.83	11.55	10.99	13.64	11.72
		w/ SRR	**22.87**	**19.54**	**25.49**	**19.98**	**11.32**	**10.75**	**12.83**	**11.45**
		QERA-exact (Zhang et al., 2025)	21.66	19.23	**22.46**	19.36	11.15	10.68	12.05	11.00
		w/ SRR	**21.24**	**18.70**	22.82	**18.65**	**11.05**	**10.53**	**11.90**	**10.74**

Table 2. Fine-tuning results on the GLUE benchmark using RoBERTa-base with PEFT under 4-, 3-, and 2-bit quantization (MXINT, blocksize 16/32). LoftQ and LQ-LoRA are run for 5 iterations. See Appendix A.2 for setup; best results are shown in **bold**.

	Method	**Rank**	**MNLI**	**QNLI**	**RTE**	**SST**	**MRPC**	**CoLA**	**QQP**	**STSB**	**Avg.**
			Acc.	Acc.	Acc.	Acc.	Acc.	Matt.	Acc.	P/S Corr.	
16	Full FT	–	87.62	93.03	76.53	95.18	89.95	61.79	91.55	90.28/90.05	85.73
	LoRA (Hu et al., 2022)	8	87.59	92.68	72.76	95.07	89.76	61.08	90.95	90.09/89.84	84.92
4.25	QLoRA (Dettmers et al., 2023)	8	86.91	92.29	66.06	94.15	86.76	56.24	90.45	88.95/88.82	82.72
	LoftQ (Li et al., 2023)		**87.13**	91.63	64.26	93.46	87.75	59.07	90.46	88.95/88.84	82.83
	QERA (Zhang et al., 2025)		87.07	92.20	64.98	94.15	87.99	58.55	90.45	89.86/89.68	83.14
	LQ-LoRA (Guo et al., 2024)		85.89	90.96	54.15	92.32	82.35	42.60	88.67	85.89/85.73	77.84
	SRR		87.09	**92.64**	**72.20**	**94.84**	**88.48**	**60.58**	**90.48**	**90.06/89.77**	**84.53**
3.25	QLoRA (Dettmers et al., 2023)	8	86.14	90.76	54.87	90.83	78.92	10.83	89.91	86.77/86.28	73.60
	LoftQ (Li et al., 2023)		86.38	90.24	57.04	91.63	81.13	14.52	89.27	86.55/86.24	74.58
	QERA (Zhang et al., 2025)		**86.49**	89.46	57.40	91.74	84.56	28.98	89.26	**87.90/87.61**	76.95
	LQ-LoRA (Guo et al., 2024)		84.70	88.74	54.51	91.63	74.75	24.37	87.61	85.16/85.31	73.95
	SRR		86.06	**91.87**	**59.93**	**93.46**	**87.50**	**50.11**	**90.01**	87.97/87.50	**80.84**
2.50	QLoRA (Dettmers et al., 2023)	64	78.58	85.34	50.98	89.22	68.63	0	88.08	66.14/66.35	65.88
	LoftQ (Li et al., 2023)		81.30	86.63	50.37	91.06	71.08	0	88.48	82.63/82.85	68.96
	QERA (Zhang et al., 2025)		84.24	88.61	54.25	90.83	81.37	21.93	89.48	83.61/83.51	74.28
	LQ-LoRA (Guo et al., 2024)		83.33	87.26	52.71	89.79	71.83	0	88.32	78.45/79.39	69.02
	SRR		**85.64**	**90.96**	**59.57**	**92.89**	**85.78**	**38.22**	**90.24**	**87.43/87.13**	**78.82**

Table 3. GSM8K results for LLaMA-2 7B fine-tuned with PEFT under 4-/2-bit quantization using MXINT (blocksize 16/32, rank 64), LoftQ and LQ-LoRA is run with 5 iterations. **Bold** indicates the highest accuracy, and detailed setup is in Appendix A.2.

		Method	**Rank**	**LLaMA-2 7B (Δ_{acc})**
Quantization Bits	**16**	LoRA (Hu et al., 2022)	64	35.41
	4.25	QLoRA (Dettmers et al., 2023)	64	32.21
		LoftQ (Li et al., 2023)		28.35
		QERA (Zhang et al., 2025)		32.13
		LQ-LoRA (Guo et al., 2024)		29.82
		SRR		**32.87**
	2.50	QLoRA (Dettmers et al., 2023)	64	14.03
		LoftQ (Li et al., 2023)		15.69
		QERA (Zhang et al., 2025)		18.76
		LQ-LoRA (Guo et al., 2024)		16.67
		SRR		**18.95**

formation in **LR**, fine-tuning all directions can degrade performance by overwriting this initialization. As shown in Table 4, partial-freezing improves performance across tasks, confirming the benefit of preserving dominant directions during adaptation. We also observe that partial-freezing allows for a larger learning rate without instability, while non-freezing requires careful tuning and performs poorly at high learning rates. Even when trained for more epochs, the non-freeze variant underperforms the partial-freeze setup.

Table 4. Comparison of adapter fine-tuning strategies with and without partial-freezing on GLUE. "Non-Freeze" updates the full adapter, while "Partial-Freeze" freezes a subset of dimensions. Best results are in **bold**, and per-task results are in Appendix C.

	Method	**Rank**	**LR**	**Epochs**	**Avg.**
4.25	SRR (Non-Freeze)	8	6e-4	5	39.16
	SRR (Non-Freeze)		7e-5	5	80.49
	SRR (Non-Freeze)		7e-5	20	83.58
	SRR (Partial-Freeze)		6e-4	5	**84.53**

5. Conclusion

We introduced *Structured Residual Reconstruction (SRR)*, a simple and effective reformulation of QER that preserves dominant directions before quantization and applies quantization only to the residual. This improves approximation quality under fixed rank constraints and generalizes to activation-aware settings and QPEFT by freezing the preserved subspace during fine-tuning. Extensive experiments across PTQ and QPEFT show that SRR consistently outperforms existing methods, offering a robust framework for quantization-aware model compression.

Impact Statement

This work advances the efficiency and reliability of deploying large language models in resource-constrained environments. By reformulating quantization error reconstruction to explicitly preserve dominant structure before quantization, our method improves both compression quality and fine-tuning stability without increasing model size. These contributions can benefit practitioners seeking to reduce the memory and compute costs of LLMs while maintaining high accuracy, thereby broadening access to state-of-the-art models in practical, real-world settings such as mobile devices, edge computing, and low-resource platforms.

References

Agirre, E., M‘arquez, L., and Wicentowski, R. (eds.). *SemEval-2007*. ACL, 2007.

Banner, R., Nahshan, Y., and Soudry, D. Post training 4-bit quantization of convolutional networks for rapid-deployment. In *Advances in Neural Information Processing Systems*, 2019.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Dagan, I., Glickman, O., and Magnini, B. The PASCAL recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*. 2006.

Darvish Rouhani, B., Zhao, R., Elango, V., Shafipour, R., Hall, M., Mesmakhosroshahi, M., More, A., Melnick, L., Golub, M., Varatkar, G., et al. With shared microexponents, a little shifting goes a long way. In *ISCA*, 2023.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. QLoRA: Efficient finetuning of quantized LLMs. In *NeurIPS*, 2023.

Dolan, W. B. and Brockett, C. Automatically constructing a corpus of sentential paraphrases. In *IWP*, 2005.

Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. OPTQ: Accurate quantization for generative pre-trained transformers. In *ICLR*, 2023.

Gao, L., Tow, J., Abbasi, B., Biderman, S., Black, S., DiPofi, A., Foster, C., Golding, L., Hsu, J., Le Noac'h, A., Li, H., McDonell, K., Muennighoff, N., Ociepa, C., Phang, J., Reynolds, L., Schoelkopf, H., Skowron, A., Sutawika, L., Tang, E., Thite, A., Wang, B., Wang, K., and Zou, A. The language model evaluation harness, 2024.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Guo, H., Greengard, P., Xing, E., and Kim, Y. Lq-lora: Low-rank plus quantized matrix decomposition for efficient language model finetuning. In *The Twelfth International Conference on Learning Representations*, 2024.

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Li, Y., Yu, Y., Liang, C., Karampatziakis, N., He, P., Chen, W., and Zhao, T. Loftq: Lora-fine-tuning-aware quantization for large language models. In *ICLR*, 2023.

Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.-M., Wang, W.-C., Xiao, G., Dang, X., Gan, C., and Han, S. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. In *MLSys*, 2024.

Liu, S.-Y., Khadkevich, M., Fung, N. C., Sakr, C., Yang, C.-H. H., Wang, C.-Y., Muralidharan, S., Yin, H., Cheng, K.-T., Kautz, J., et al. Eora: Training-free compensation for compressed llm with eigenspace low-rank approximation. *arXiv preprint arXiv:2410.21271*, 2024.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer sentinel mixture models. In *ICLR*, 2017.

Nagel, M., Amjad, R. A., Van Baalen, M., Louizos, C., and Blankevoort, T. Up or down? adaptive rounding for post-training quantization. In *ICML*, 2020.

Nagel, M., Fournarakis, M., Amjad, R. A., Bondarenko, Y., Van Baalen, M., and Blankevoort, T. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*, 2021.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. SQuAD: 100,000+ questions for machine comprehension of text. In *EMNLP*, 2016.

Rivière, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., et al. Gemma 2: Improving open language models at a practical size. In *CoRR*, 2024.

Saha, R., Sagan, N., Srivastava, V., Goldsmith, A., and Pilanci, M. Compressing large language models using low rank and low precision decomposition. In *NeurIPS*, 2024.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 2013.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Tseng, A., Chee, J., Sun, Q., Kuleshov, V., and De Sa, C. Quip#: Even better llm quantization with hadamard incoherence and lattice codebooks. In *ICML*, 2024.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*, 2019.

Warstadt, A., Singh, A., and Bowman, S. Neural network acceptability judgments. In *TACL*, 2019.

Williams, A., Nangia, N., and Bowman, S. R. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*, 2018.

Yao, Z., Yazdani Aminabadi, R., Zhang, M., Wu, X., Li, C., and He, Y. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. In *NeurIPS*, 2022.

Yao, Z., Wu, X., Li, C., Youn, S., and He, Y. Exploring post-training quantization in llms from comprehensive study to low rank compensation. In *AAAI*, 2024.

Zhang, C., Cheng, J., Constantinides, G. A., and Zhao, Y. Lqer: low-rank quantization error reconstruction for llms. In *ICML*, 2024a.

Zhang, C., Wong, J. T. H., Xiao, C., Constantinides, G. A., and Zhao, Y. QERA: an analytical framework for quantization error reconstruction. In *ICLR*, 2025.

Zhang, P., Zeng, G., Wang, T., and Lu, W. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*, 2024b.

A. Experiment Details

A.1. Experiment Details for PTQ

Setup and Quantization Configuration. We evaluate our method under the post-training quantization (PTQ) paradigm across various scales: TinyLlama-1.1B (Zhang et al., 2024a), Gemma-2 2B (Rivière et al., 2024), LLaMA-2 7B (Touvron et al., 2023), and LLaMA-3.1 8B (Grattafiori et al., 2024). Weights are quantized to both 4-bit and 3-bit precision using MXINT (Darvish Rouhani et al., 2023) with block size 32, yielding effective bitwidths of 4.25 and 3.25, respectively. Low-rank correction terms are computed with rank 32 for 4-bit and rank 64 for 3-bit quantization.

Evaluation Benchmarks. To assess performance comprehensively, we report both language modeling perplexity and downstream task accuracy. Perplexity is evaluated on WikiText2 (Merity et al., 2017) using lm-evaluation-harness (Gao et al., 2024).

Baselines. We benchmark our approach against leading PTQ baselines that incorporate low-rank quantization error reconstruction, including ZeroQuant-V2 (Yao et al., 2024), LQER (Zhang et al., 2024a), and QERA (Zhang et al., 2025) in both its approximate and exact forms. All baseline implementations are standardized to use the same quantization format, block size, and calibration data to ensure fair comparison. In addition, we include quantization-only models (*w-only*) to isolate the effect of low-rank correction.

A.2. Experiment Details for QPEFT

Configuration Before Finetuning. Baselines include QLoRA (Dettmers et al., 2023), LoftQ (Li et al., 2023), QERA (Zhang et al., 2025), and LQ-LoRA (Guo et al., 2024). We employ MXINT (Darvish Rouhani et al., 2023), a block-wise quantization method that leverages shared codebooks to enhance compression efficiency while preserving model fidelity. GLUE tasks are quantized to 4- and 3-bit precision with block size 32 and PEFT rank 8, while 2-bit quantization uses block size 16 with rank 64. For GSM8K, we use both 2-bit (block size 16) and 4-bit (block size 32) configurations, each with a PEFT rank of 64. Additionally, we adopt five iterations for LoftQ (Li et al., 2023) and LQ-LoRA (Guo et al., 2024). For QERA (Zhang et al., 2025), we consistently adopt the exact scaling mode, as its second-order activation statistics can be computed once and reused across PTQ and QPEFT stages, ensuring consistency. While LQ-LoRA originally applies a Fisher-weighted objective to determine scaling directions, we instead use the same exact scaling for all methods—including LQ-LoRA—for fair comparison. This choice provides a unified activation-aware basis across baselines and eliminates confounding factors arising from inconsistent scaling heuristics.

Fine-tuning on Natural Understanding Tasks: GLUE. We evaluate our approach on the GLUE benchmark (Wang et al., 2019), which comprises eight diverse tasks: MNLI (Williams et al., 2018), QNLI (Rajpurkar et al., 2016), RTE (Dagan et al., 2006), SST-2 (Socher et al., 2013), MRPC (Dolan & Brockett, 2005), QQP, CoLA (Warstadt et al., 2019), and STSB (Agirre et al., 2007), and task descriptions are summarized in Table 5. Following prior work (Zhang et al., 2025), we report accuracy for MNLI, QNLI, RTE, SST-2, MRPC, and QQP; Matthews correlation for CoLA; and Pearson/Spearman correlations for STSB. All methods are built upon RoBERTa-base (Liu et al., 2019) and fine-tuned using a consistent strategy, where each model is trained for five epochs. To ensure fair comparisons, task-specific learning rates are selected, with details provided in Table 6.

Table 5. Descriptions of the eight GLUE tasks used in our evaluation.

Task	Description
MNLI	Infer relation: entailment / neutral / contradiction
QNLI	Does the context sentence answer the question?
RTE	Does the premise entail the hypothesis?
SST-2	Sentiment classification (positive/negative)
MRPC	Are the two sentences paraphrases?
CoLA	Grammatical acceptability
QQP	Are the two questions semantically equivalent?
STSB	Predict semantic similarity score (0–5) between sentences

Table 6. Learning rates of RoBERTa-base experiments on GLUE.

Bits	**Rank**	**Method**	**Learning Rates**
16	–	Full FT	7e-5, 5e-5, 3e-5, 2e-5
16	8	LoRA	3e-4, 5e-4, 6e-4, 7e-4
4.25	8	QLoRA / LoftQ / QERA / SRR	3e-4, 5e-4, 6e-4, 7e-4
4.25	8	LQ-LoRA	5e-5, 7e-5, 1e-4, 6e-4
3.25	8	QLoRA / LoftQ / QERA / SRR	1e-4, 3e-4, 5e-4, 6e-4
3.25	8	LQ-LoRA	3e-5, 5e-5, 7e-5, 5e-4
2.50	64	QLoRA / LoftQ / QERA / SRR	5e-5, 7e-5, 9e-5, 1e-4, 2e-4
2.50	64	LQ-LoRA	1e-5, 3e-5, 5e-5, 7e-5, 9e-5

Fine-tuning on Reasoning and Language Modeling Tasks: GSM8K. To assess generative capabilities, we evaluate on GSM8K (Cobbe et al., 2021) for arithmetic reasoning, which is framed as causal language modeling (CLM), with evaluation based on exact-match accuracy. We fine-tune LLaMA-2 7B (Touvron et al., 2023) using PEFT adapters under 4- and 2-bit quantization regimes. Models are trained for 10 epochs with a total batch size of 32, and learning rates are swept over $7\text{e}{-5}, 1\text{e}{-4}, 3\text{e}{-4}, 5\text{e}{-4}$, excluding LQ-LoRA, for which learning rates are swept over $3\text{e}{-5}, 5\text{e}{-5}, 7\text{e}{-5}, 1\text{e}{-4}$. The best-performing configuration for each method is reported.

A.3. Other Settings

Here, we list the models used in our experiments, along with their sources and licensing information, respectively in Table 7. Besides, experiments were executed using both eight NVIDIA A100 and eight NVIDIA L40S GPUs, distributed across separate machines. GPU usage was adjusted based on the task: GLUE experiments were performed on a single GPU per run, whereas GSM8K experiments utilized 4 GPUs concurrently. All reported results are averaged over three runs with different random seeds to ensure robustness.

Table 7. Summary of models used in this paper, including source, access method, and license.

Model	**Source**	**Accessed via**	**License**
RoBERTa-Base	(Liu et al., 2019)	Link	MIT License
TinyLlama-1.1B	(Zhang et al., 2024b)	Link	Apache License 2.0
Gemma-2-2B	(Rivière et al., 2024)	Link	Gemma License
LLaMA-2-7B-hf	(Touvron et al., 2023)	Link	LLaMA 2 Community License
LLaMA-2-13B	(Touvron et al., 2023)	Link	LLaMA 2 Community License
LLaMA-3.1-8B	(Grattafiori et al., 2024)	Link	LLaMA 3.1 Community License

B. Additional PTQ Experiments and Analysis

Performance under 4-bit Quantization. Table 8 demonstrates that applying SRR (w/ SRR) reliably reduces perplexity across all models and base quantization methods under 4-bit PTQ. Improvements are stable across both rank settings, indicating that SRR effectively decomposes and preserves activation-sensitive structure during quantization. This supports our central claim that preserving dominant directions before quantization leads to more robust and accurate low-rank recovery.

Performance on ZeroQuant-V2. SRR does not always outperform ZeroQuant-V2 in perplexity, even though it consistently reduces the weight reconstruction error. As shown in Figure 4, SRR achieves lower $\|\mathbf{W} - \mathbf{Q} - \mathbf{LR}\|_F$ across all layers compared to existing QER frameworks, confirming its effectiveness in approximating the original weights. However, in cases where SRR underperforms, the root cause lies in the limitations of ZeroQuant-V2, which does not account for activation statistics during quantization. As a result, minimizing the weight error does not necessarily lead to reduced output error when passing through the quantized layer. Therefore, the observed performance gap is not a failure of SRR, but rather a consequence of applying accurate weight reconstruction within a pipeline that lacks activation-aware calibration.

SRR in Iterative Quantization. To evaluate our method in iterative quantization settings, we apply SRR in an alternating update scheme, similar to LoftQ (Li et al., 2023), LQ-LoRA (Zhang et al., 2024a), and CALDERA (Liu et al., 2024). These

Table 8. Perplexity (↓) of quantized LLMs on WikiText2 under 4-bit PTQ using MXINT. We apply our method to various existing baselines across four models and two low-rank settings (**LR** rank $r = 8$ and $r = 32$). Lower perplexity scores are highlighted in **bold**.

		Method	**TinyLlama 1.1B**		**Gemma-2 2B**		**LLaMA-2 7B**		**LLaMA-3.1 8B**	
			$r = 8$	$r = 32$	$r = 8$	$r = 32$	$r = 8$	$r = 32$	$r = 8$	$r = 32$
		BF16	13.98		13.08		8.71		7.55	
Quantization Bits	**4.25**	*w-only*	19.40		16.23		9.45		8.78	
		ZeroQuant-V2	16.98	16.60	16.01	15.60	**9.42**	**9.42**	**8.87**	**8.83**
		w/ SRR	**15.80**	**16.51**	**15.82**	**15.48**	9.50	9.45	8.92	8.92
		LQER	16.31	15.83	15.30	14.46	**9.27**	**9.22**	**8.55**	**8.46**
		w/ SRR	**15.46**	**15.24**	**15.11**	**14.10**	**9.27**	**9.22**	**8.55**	8.47
		QERA-approx	16.01	15.39	**15.12**	14.49	9.23	9.17	8.53	8.45
		w/ SRR	**15.45**	**15.28**	**15.12**	**14.28**	**9.19**	**9.13**	**8.52**	**8.43**
		QERA-exact	15.27	15.63	**14.50**	14.26	9.17	9.12	8.42	8.33
		w/ SRR	**15.16**	**15.01**	14.66	**14.20**	**9.16**	**9.09**	**8.41**	**8.32**

approaches perform *iterative updates that alternate between quantizing weights and computing low-rank SVDs*, enabling mutual refinement of quantized and reconstructed components. To remain faithful to the design of SRR—which aims to preserve dominant directions in the final model—we ensure that the last step always applies low-rank reconstruction rather than quantization.

Table 9 and Table 10 compare three SRR configurations:

- **None**: No dominant direction is preserved. This reduces SRR to a plain QER setup, where low-rank parameters are trained after quantization without any structural guidance.
- **Full**: Low-rank reconstruction is performed *before* quantization, using the full set of r directions. Since no selection is applied, this configuration attempts to reshape the entire weight matrix in advance.
- **Top-k (Ours)**: Only the most impactful k directions, ranked by score, are retained. By focusing reconstruction on this compact subspace, our method preserves essential information while minimizing interference with quantization.

Across both 3-bit and 4-bit settings, the Top-k configuration consistently achieves the best or near-best perplexity, outperforming both the unstructured (Full) and naive (None) alternatives. These results highlight the effectiveness of our rank selection strategy and further suggest that SRR provides a strong initialization for iterative quantization, helping guide the optimization toward better local minima from the outset.

Table 9. Perplexity (↓) on WikiText2 under 3-bit PTQ using MXINT, measured after 5 iterative reconstruction steps. We compare three SRR configurations: None (no preservation), Full (all directions preserved before quantization), and Top-k (ours), which retains only the top-ranked directions. Lower scores are highlighted in **bold**.

	SRR Rank Selection	**TinyLlama 1.1B**		**Gemma-2 2B**		**LLaMA-2 7B**		**LLaMA-3.1 8B**	
		$r = 8$	$r = 64$	$r = 8$	$r = 64$	$r = 8$	$r = 64$	$r = 8$	$r = 64$
3.25	None	21.10	18.22	**21.70**	17.96	11.05	10.48	12.24	10.60
	Full	21.00	17.82	21.91	17.48	**10.95**	10.41	11.80	10.40
	Top-k (Ours)	**20.87**	**17.77**	21.90	**17.33**	11.02	**10.37**	**11.75**	**10.39**

Table 10. Perplexity (↓) on WikiText2 under 4-bit PTQ using MXINT, measured after 5 iterative reconstruction steps. We compare three SRR configurations: None (no preservation), Full (all directions preserved before quantization), and Top-k (ours), which retains only the top-ranked directions. Lower scores are highlighted in **bold**.

	SRR Rank Selection	**TinyLlama 1.1B**		**Gemma-2 2B**		**LLaMA-2 7B**		**LLaMA-3.1 8B**	
		$r = 8$	$r = 32$	$r = 8$	$r = 32$	$r = 8$	$r = 32$	$r = 8$	$r = 32$
4.25	None	15.19	14.97	**14.47**	**14.07**	9.20	9.10	8.42	8.28
	Full	15.14	14.93	14.49	14.08	9.17	**9.07**	**8.38**	8.26
	Top-k (Ours)	**15.12**	**14.85**	14.53	**14.07**	**9.15**	**9.07**	8.39	**8.25**

C. Discussion on QPEFT

Full results on importance of freezing dominant directions. As shown in Table 11, partial freezing consistently outperforms full updates under the same rank. This confirms that preserving dominant directions leads to better fine-tuning stability and accuracy.

Table 11. Comparison of adapter fine-tuning strategies with and without partial-freezing on the GLUE benchmark (RoBERTa-base). "Non-Freeze" updates the entire adapter, while "Partial-Freeze" selectively freezes a subset of adapter dimensions. Best results are in **bold**.

	Method	**Rank**	**LR**	**Epochs**	**MNLI**	**QNLI**	**RTE**	**SST**	**MRPC**	**CoLA**	**QQP**	**STSB**	**Avg.**
					Acc.	Acc.	Acc.	Acc.	Acc.	Matt.	Acc.	P/S Corr.	
4.25	SRR (Non-Freeze)	8	6e-4	5	32.95	50.54	47.29	50.92	68.38	0	63.18	NaN/NaN	39.16
	SRR (Non-Freeze)		7e-5	5	86.13	90.99	63.90	90.58	87.75	48.75	88.82	87.00/87.07	80.49
	SRR (Non-Freeze)		7e-5	20	86.81	92.28	69.68	93.23	88.73	57.83	**90.48**	89.72/89.50	83.58
	SRR (Partial-Freeze)		6e-4	5	**87.09**	**92.64**	**72.20**	**94.84**	**88.48**	**60.58**	**90.48**	**90.06/89.77**	84.53

D. Additional Results on SRR error distribution

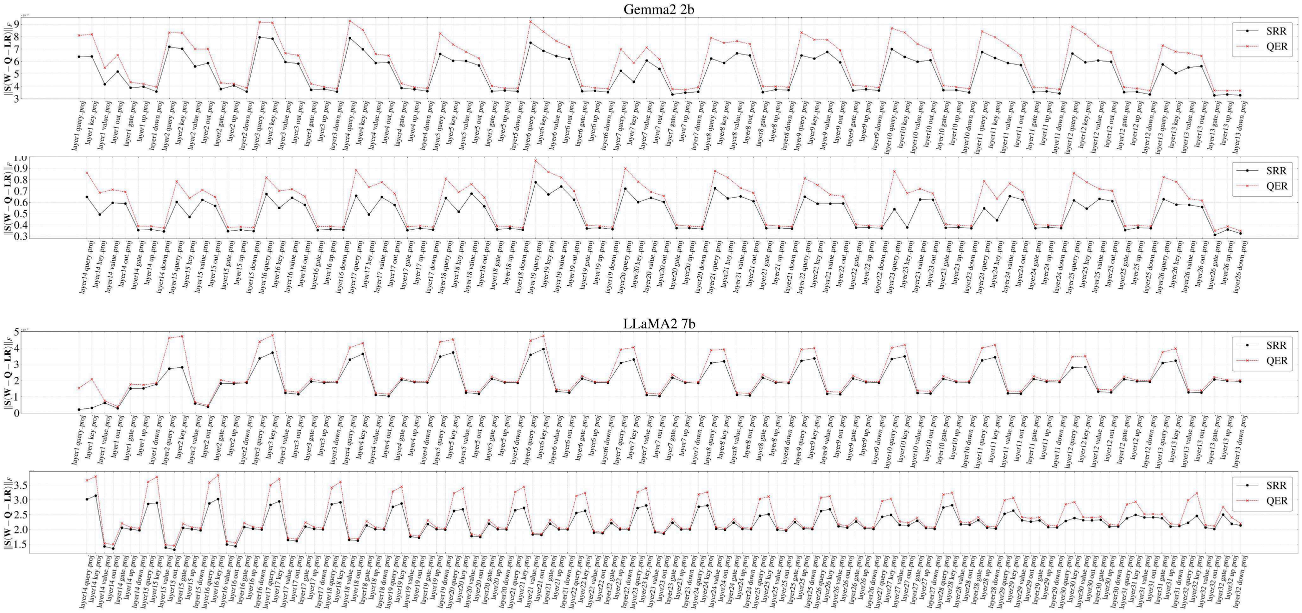

Figure 4. Layer-wise decomposition error (full) using SRR and QER under $\mathbf{S} = \mathbf{I}$ (identity). SRR consistently achieves lower error than QER in both Gemma-2 (top) and LLaMA-2 (bottom).

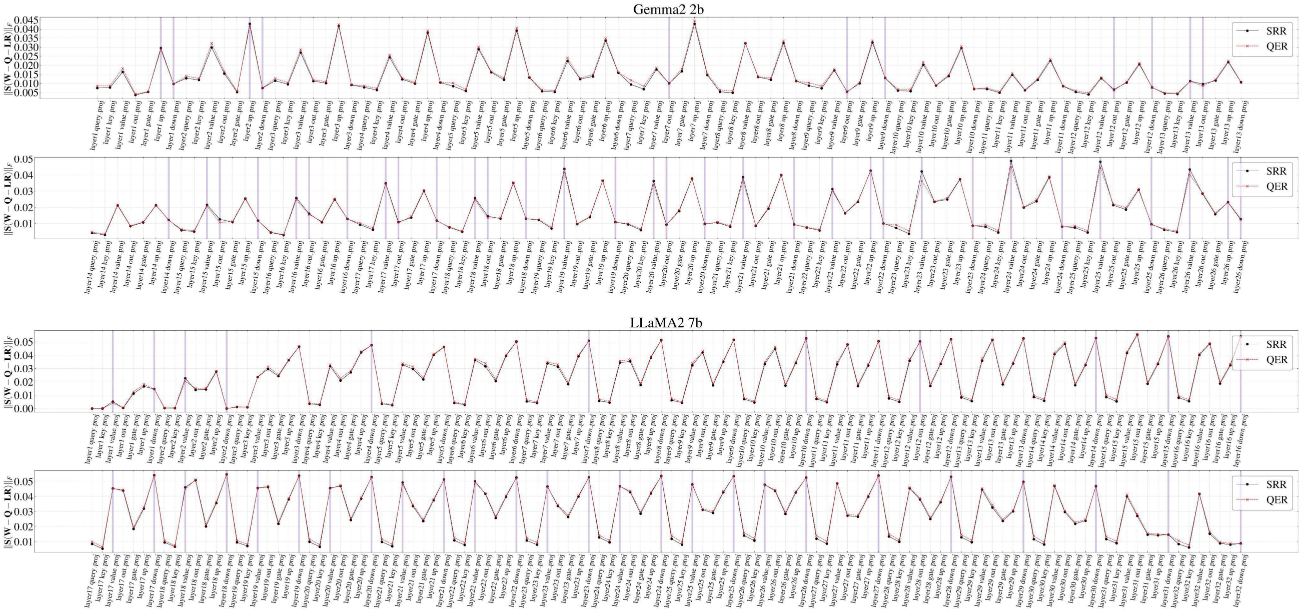

Figure 5. Layer-wise decomposition error (Full) of SRR and QER under $\mathbf{S} \neq \mathbf{I}$ (QERA-exact). SRR does not always outperform QER in either Gemma-2 (top) or LLaMA-2 (bottom). Layers where SRR performs worse are shown in purple.