GRACE-SVD: Guided Rank Allocation and Core Equalization SVD

Andra Siva Sai Teja* **Pragya Paramita Sahu*** **Vibhor Dave***
Joshua Mahadevan **Ankit Mahato** **Debanjan Konar** **Vikram Mupparthi**
Samsung R&D Institute India, Bangalore
{andra.teja, pragya.sahu, vibhor.dave}@samsung.com

Abstract

Large Language Models (LLMs), with 7B+ parameters, remain difficult to deploy on edge devices due to memory & latency constraints, despite increasing demand for personalized on-device agents with strong reasoning capabilities. Enabling such deployment, in current hardware specifications, requires aggressive compression while preserving downstream performance. Singular Value Decomposition (SVD)-based methods provide a promising reconstruction-aware compression paradigm, and recent trainable frameworks improve adaptive rank allocation under varying compression budgets. However, maintaining model quality under high compression remains challenging due to statistical mismatches between decomposed cores and compounded degradation from direct high-ratio compression.

We propose **GRACE-SVD**, a trainable SVD compression framework combining *Guided Rank Allocation* with a *novel Core Equalization* method that aligns decomposed core distributions while preserving reconstruction. We further introduce a progressive compression pipeline with intermediate equalization to mitigate performance collapse under aggressive compression. Across multiple model families and compression settings, our GRACE-SVD consistently outperforms existing SVD techniques, achieving stable compression even at 0.4 parameter ratio, with lower perplexity and stronger zero-shot performance under comparable or lower compression budgets.

1 Introduction

Large Language Models (LLMs) have transformed modern AI, enabling applications ranging from language understanding to multimodal reasoning. As the reasoning engine behind AI agents, LLMs are increasingly expected to perform personalized assistance and on-device decision making. However, deploying such capabilities directly on edge devices remains challenging due to the substantial memory and compute requirements of modern models. While parameter-efficient techniques like low-rank adapters (LoRA) (Hu et al., 2021; Zhang et al., 2023; Aastik et al., 2025) are widely used to efficiently enable a single base model to perform multiple downstream tasks, the deployment footprint of the base model remains large. Cloud-based inference introduces significant privacy and latency concerns, particularly when handling sensitive user data in local settings such as on-device visual question answering (Bara et al., 2022; Sahu et al., 2024). Efficient compression of large models has therefore become a critical area of research.

Techniques like Quantization (Frantar et al., 2022; Lin et al., 2024), Pruning (Yang et al., 2025b; Sun et al., 2024) and Distillation (Sanh et al., 2019; Tan et al., 2023) have been widely studied, achieving substantial reductions in memory and inference cost. However, these techniques often suffer noticeable degradation under aggressive compression settings or require specialized hardware support for efficient deployment. In contrast, Singular Value Decomposition (SVD) based compression techniques have opened up a new paradigm of exploration, providing a mathematically stable method of approximating the same information with lesser parameters. Of late, *Dobi-SVD : Differentiable SVD for LLM Compression and Some New Perspectives* (Wang et al., 2025a) and *ARA : Adaptive Rank Allocation for Efficient Large Language Model SVD Compression* (Xv et al., 2026) address the layer sensitivity problem with a trainable optimization framework for layer-wise rank allocation. Despite these advances, challenges related to stability during differentiable optimization, robust rank allocation, and downstream adaptation under high compression remain unresolved.

*Equal contribution, with the order of authors decided on the basis of alphabetical order

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 45569–45587
October 24-29, 2026 ©2026 Association for Computational Linguistics

In this paper, we propose GRACE-SVD, a trainable SVD-based compression framework for efficient LLM decomposition under high compression settings. Our primary contributions are:

1. We introduce a stable training framework and adaptive rank selection pipeline capable of identifying effective low-rank, layer-wise representations under parameter ratios as low as 0.4, while maintaining minimal perplexity degradation and less than 5% reduction in downstream task performance.
2. We identify the statistical distribution mismatches between decomposed low-rank cores (which was causing significant challenges during downstream adaptation) and propose a novel core distribution equalization framework that improves decomposition stability.
3. We extensively evaluate both direct and progressive compression strategies across multiple model families and scales, demonstrating consistent improvements over existing SVD-based compression approaches under high compression regimes.

These findings highlight the importance of adaptive decomposition and distribution-aware optimization in enabling reliable high-ratio compression. Core equalization codes are released here [1].

2 Related Work

While LLMs have redefined AI capabilities, its wide-spread deployment has been constrained by compute and memory limitations. Compression techniques have been explored to address this issue.

Model Pruning removes redundant parameters to reduce model complexity. Unstructured pruning methods (Frantar and Alistarh, 2023; Sun et al., 2024; Yang et al., 2025b) achieve high sparsity, while Deja Vu (Liu et al., 2023) exploits contextual sparsity and SEAP (Liang et al., 2025) identifies task-specific activation paths. But these high sparsity numbers do not consistently translate to improved latency because of hardware limitations. Conversely, structured pruning (Ma et al., 2023; Xia et al., 2024; Ashkboos et al., 2024) removes entire model components, enabling more hardware-friendly acceleration. But they typically suffer from significant loss at high compression rates.

Quantization methods have tackled the challenge of memory footprint minimization using bit-precision reduction. Foundational methods include GPTQ (Frantar et al., 2022), SmoothQuant (Xiao et al., 2023), and AWQ (Lin et al., 2024) which improve quantization robustness through second-order optimization, activation smoothing, and salient weight preservation, respectively. More recently, ultra-low precision methods such as 1-bit LLM quantization have also been explored (Xu et al., 2024). Despite their efficiency gains, quantization methods often face accuracy degradation under aggressive low-bit compression settings.

Knowledge Distillation techniques train a smaller 'student' model to mimic the behaviour of a large 'teacher' model, using a combination of losses to align output probabilities & hidden states. DistilBERT (Sanh et al., 2019) introduced an effective distillation framework, achieving substantial reductions in model size while retaining most of BERT's performance. Subsequent works such as TinyBERT (Jiao and et al., 2020), GKD (Tan et al., 2023), and ResidualKD (On et al., 2026) further improve representation alignment & distillation objectives, enabling compact student models with highly competitive performance. Such similar distillation frameworks have also been successfully applied in computer vision to create significantly smaller models while preserving performance (Jha et al., 2024).

Singular Value Decomposition based low-rank compression methods have been an important compression technique, reducing model size by factorizing weight matrices into low rank representations. ASVD (Yuan and et al., 2024) introduced activation-aware scaling to preserve important weights, while SVD-LLM (Wang et al., 2025c) proposed an optimal scaling strategy which improved LLM performance. Recent works such as ARA(Xv et al., 2026) shifted SVD-based compression from static ranks toward fully trainable optimization frameworks, enabling adaptive layer-wise rank allocation guided by downstream performance. More recently, Swift-SVD (Qi et al., 2026) proposed a fast, training-free, and numerically stable formulation, significantly reducing compression overhead.

Recent advances in trainable SVD compression highlight the growing importance of adaptive optimization strategies for maintaining model quality under high compression. Our work builds

[1] https://github.com/SamsungLabs/LACE

on this direction by combining differentiable rank allocation with distribution-aware core equalization for stable and efficient SVD-based LLM compression. Specifically, we target aggresive compression using optimized decomposition methods, that can be utilised in tandem with downstream binarization and quantization techniques to attain close-to-base performance at a fraction of memory cost.

3 Methods

In this work, we propose **GRACE-SVD : Guided Rank Allocation & Core Equalization - Singular Value Decomposition** method, to compress LLMs through efficient low-rank decomposition. We identify inefficiencies in current approaches and propose novel approaches to address them, through our proposed two-staged setup:

- **Data-*G*uided SVD *R*ank *A*llocation** : We build upon the works of (Xv et al., 2026) to build a trainable pipeline, that determines optimal rank per layer based on its impact on model performance (measured through a weighted combination of compression and task losses). This ensures that for each target compression, we are able to attain the most efficient core-decomposition of each layer - allowing the framework to generalize across varying model scales and architectures. We also modify the regular SVD algorithm, as discussed in Section 3.1, to improve overall pipeline stability and performance.
- ***C*ore *E*qualized Weight Re-mapping**: We notice a significant statistical dissimilarity between the decomposed cores, that adversely affects downstream learning capabilities, across multiple PEFT and SFT techniques. We propose a statistical remapping followed by short-horizon logit alignment for the decomposed cores to improve learning capabilities of our final decomposed model in downstream tuning.

We evaluate both direct & progressive applications of the proposed pipeline under high compression settings, and discuss their comparative behavior and implications in Section 4.3.2. The following sections describe the proposed techniques in detail.

3.1 Stable SVD Framework

Consider a linear layer with weight matrix $W \in \mathbb{R}^{m\times n}$. Our goal is to decompose W into two lower-rank matrices, $A \in \mathbb{R}^{m\times r}$ and $B \in \mathbb{R}^{r\times n}$, such that $r \ll \min(m,n)$, achieved by preserving only the **top-r singular values**. This reduces the parameter count from mn to $r(m+n)$ while approximating the original transformation through $W' = AB$. Optimal loss reduction is obtained by minimization of:

$$\min_{W'} \mathcal{L}_r = \|WX - W'X\|_F \quad \text{s.t.} \quad \text{rank}(W') \le r \tag{1}$$

where $X \in \mathbb{R}^{n\times d}$ denotes the layer input. Following (Wang et al., 2025c), we use an activation-aware SVD formulation, involving a Cholesky transformation of the input activations to obtain the whitening matrix $S \in \mathbb{R}^{n\times n}$. The decomposed matrices are thus obtained as:

$$\begin{aligned} U\Sigma V^T &= SVD(WS) \\ A = U\sqrt{\Sigma_r}, &\qquad B = \sqrt{\Sigma_r}V^T S^{-1} \end{aligned} \tag{2}$$

Σ_r contains the top-r singular values of WS.

Trainable rank-allocation methods require back-propagation through SVD, which is known to be numerically unstable when singular values are nearly identical or approach zero, causing gradient explosions during optimization (Wang et al., 2019). Equation 1 from prior work (Wang et al., 2025a) shows that the gradient with respect to the decomposed matrix depends inversely on

$$E_{ij} = \begin{cases} \sigma_j^2 - \sigma_i^2, & i \neq j \\ 1, & i = j \end{cases}$$

where σ_i, σ_j denote singular values. Consequently, gradients become unstable when $\sigma_i \approx \sigma_j$. To mitigate this issue, we adopt the stabilization strategy of (Wang et al., 2025a), replacing the unstable inverse term with a truncated Taylor-series approximation:

$$\frac{1}{E_{ij}} \approx \frac{1}{\sigma_i(\sigma_i + \sigma_j)} \sum_{p=0}^{K} \left(\frac{\sigma_j}{\sigma_i}\right)^p$$

This approximation improves numerical stability and enables reliable gradient propagation through SVD during rank optimization.

3.2 Guided Rank Allocation via Probability Simplex

For any low-rank decomposition framework, selecting the appropriate rank for each layer is critical, as reconstruction quality & downstream task performance are highly sensitive to layer-wise rank

allocation. Even under identical compression budgets, different rank distributions across layers can lead to significant variations in model behavior and accuracy. Specifically, (Xv et al., 2026) introduces a framework that utilizes a **probability simplex** to decide the best rank for a specific layer. Building on this, we develop a stable and computationally efficient pipeline to address the practical gaps in the original formulation. We utilize ***softmax-normalized logits*** to obtain the probability simplex α, enabling differentiable & interpretable rank allocation during optimization. This provides stable gradient flow, avoids premature collapse to suboptimal rank configurations and enables smoother exploration of the compression-performance trade-off.

Specifically, let $\alpha_{logits} \in \mathbb{R}^{(1\times d)}$ denote the learnable logits corresponding to candidate rank configurations. These logits are converted into a probability simplex using the softmax function:

$$\alpha = \text{Softmax}(\alpha_{logits}) = \frac{\exp(\alpha_{logits,i})}{\sum_j \exp(\alpha_{logits,j})} \quad (3)$$

ensuring 'simplex' distribution without the need to manually 'clip' numbers as shown in Fig 1. Output distribution constructs the **probability mask** $p \in \mathbb{R}^{(1\times r)}$ through a staircase matrix $M \in \mathbb{R}^{(d\times r)}$ [2]

$$p = \alpha M \quad (4)$$

We utilize this probability mask p to compute the rank of compression $r_{comp} = \sum_{i=0}^{r} p_i$ as shown in Figure 1. Thus, the compression ratio(R) for a weight matrix W, is computed as:

$$R = 1 - \frac{(r_{comp})(m+n)}{mn} \quad (5)$$

where, m and n are the weight matrix dimensions.

3.2.1 Training Objectives

Following (Xv et al., 2026), we optimize the proposed framework using a combination of three objectives: (i) Cross-Entropy loss L_{ce}, (ii) Compression loss L_{comp}, and (iii) Guidance loss L_g.

The cross-entropy loss preserves downstream language modeling performance, while the compression loss encourages lower effective rank utilization during optimization. The guidance loss acts as a teacher-student regularization objective by aligning the intermediate representations of the compressed model with those of the original model, helping preserve model behavior under aggressive compression settings.

The overall optimization objective is given by:

$$L = \lambda_{ce}L_{ce} + \lambda_{comp}L_{comp} + \lambda_g L_g \quad (6)$$

The proposed rank-allocation pipeline is trained on WikiText dataset using the combined optimization objective until convergence. Further details on individual loss components and implementation specifics are covered in Appendix A.2.

3.3 Core Equalized Weight Updates

Following the decomposition formulation in Equation 2, each weight matrix W is represented using two low-rank matrices (A and B), which we refer to as cores. Since the decomposition is performed through the input-aware transformed representation WS, the resulting cores exhibit distinct statistical characteristics after reconstruction through S^{-1} as shown in Equation 2.

Assuming the weights $w_{i,j}$ of a matrix W are sampled from an underlying distribution $f_W(\cdot)$, the corresponding core distributions can similarly be estimated from their decomposed parameters. We notice that there is a significant mismatch in the distributions $f_A(.)$ and $f_B(.)$, also visualized in Figure 2, Figure 1(b). This statistical imbalance introduces several challenges:

1. Uneven parameter scaling across the decomposed cores, leading to suboptimal reconstruction behavior and unstable layer-wise representations.

2. Reduced effectiveness of downstream adaptation methods (e.g., PEFT, SFT, and distillation), since both cores are optimized jointly despite exhibiting highly different statistical characteristics. This behaviour is illustrated in Figure 7 of Appendix.

3. Increased quantization error at module and layer granularity, particularly under aggressive low-bit compression settings.

The differing statistical behavior of the decomposed cores must be accounted for during optimization. While this could be handled through separate distribution-aware gradient updates, such approaches significantly increase training complexity. Instead, we reduce the distributional mismatch between the cores themselves while preserving the reconstructed weight representation. To this end,

[2] A structured binary matrix where successive rows activate progressively fewer (or more) columns

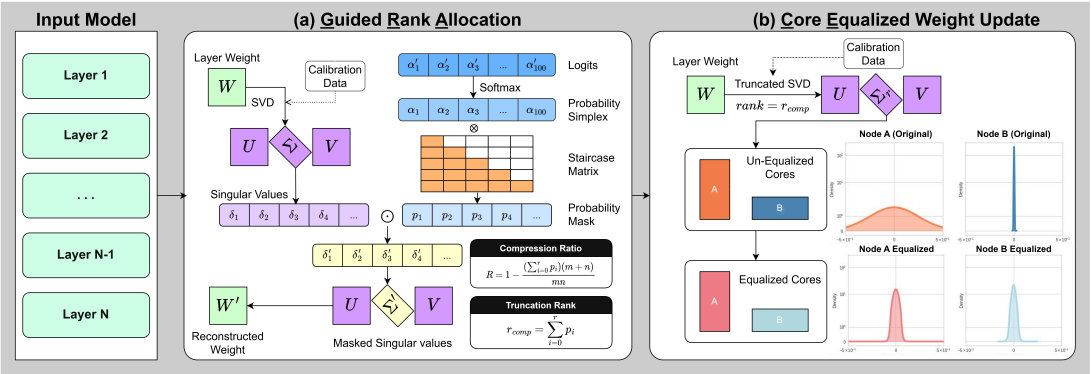


Figure 1: **GRACE-SVD: Guided Rank Allocation and Core Equalized Low-Rank Decomposition of LLMs.** (a) Data-guided rank allocation via softmax-normalized logits over stable SVD, jointly optimizing compression and task objectives for per-layer rank selection; (b) Core equalized weight updation via Wasserstein-optimal scaling to align decomposed core distributions, improving adaptation stability.

we introduce a novel ***core equalization method*** that aligns the underlying weight distributions of the decomposed cores, improving downstream adaptation stability and reducing quantization error. The core distributions are estimated using histogram-based density estimation, which provides a simple and computationally efficient approximation of the underlying statistics.

3.3.1 Core Distribution Equalization

Let $f_A(\cdot)$ and $f_B(\cdot)$ denote the distributions of the decomposed cores A and B. Let us denote A', B' as some transformations on cores A, B respectively. Our goal is to transform the cores such that $f_{A'}(\cdot) \approx f_{B'}(\cdot)$ while preserving the reconstruction constraint $A'B' = AB$.

To achieve this, we restrict the formulation to linear product-preserving transformations of the form $A' = cA$ and $B' = \frac{B}{c}$, reducing the problem to finding a factor c such that

$$f_{cA}(\cdot) \approx f_{c^{-1}B}(\cdot) \tag{7}$$

where c may depend on the statistics of the corresponding cores A and B (i.e, $c \equiv c(A, B)$) , quantifying the discrepancy between the distributions. We use the Wasserstein Distance (Panaretos and Zemel, 2019) to tackle this.

Proposition 1 ***Wasserstein-2 Distance of Two Uniform Random Variables*** *Let $X \sim \mathcal{U}(a, b)$ and $Y \sim \mathcal{U}(c, d)$. The squared Wasserstein-2 distance is:*

$$W_2^2(X, Y) = (\mu_X - \mu_Y)^2 + \frac{(R_X - R_Y)^2}{12} \tag{8}$$

where μ_X and μ_Y are the means, and R_X and R_Y are the ranges of X and Y respectively ($R_X = b - a$ and $R_Y = d - c$).

Directly aligning the raw weight distributions, however, may not accurately reflect parameter importance. As illustrated in Figure 2(a), most weights are concentrated near zero magnitude, while higher magnitude weights occur less frequently. Despite their lower frequency, such weights can remain important due to activation outliers and layer-specific sensitivities. We therefore construct an importance distribution over the weights (operational weight-range) and optimize the Wasserstein distance between the corresponding core importance distributions. For computational efficiency, we approximate the importance distributions using uniform distributions (Figure 2(b)).

Modern neural network weight distributions are typically centered around zero due to standard optimization dynamics (Glorot and Bengio, 2010; He et al., 2015). Empirically, we observe a similar near zero-mean behavior. Under this assumption, Equation 8 simplifies to

$$W_2^2(X, Y) \approx \frac{(R_X - R_Y)^2}{12} \tag{9}$$

Thus, minimizing the Wasserstein distance reduces to matching the ranges of the corresponding importance distributions. Instead of directly aligning the raw distributions of A' and B', we therefore align their importance distributions, denoted as $f^I_{A'}(\cdot)$ and $f^I_{B'}(\cdot)$, and use the resulting factor to scale the original cores.

3.3.2 Equalization Strategies

Strategy-i: Range-Based Equalization We approximate the importance distributions using the full value ranges of the decomposed cores. Let

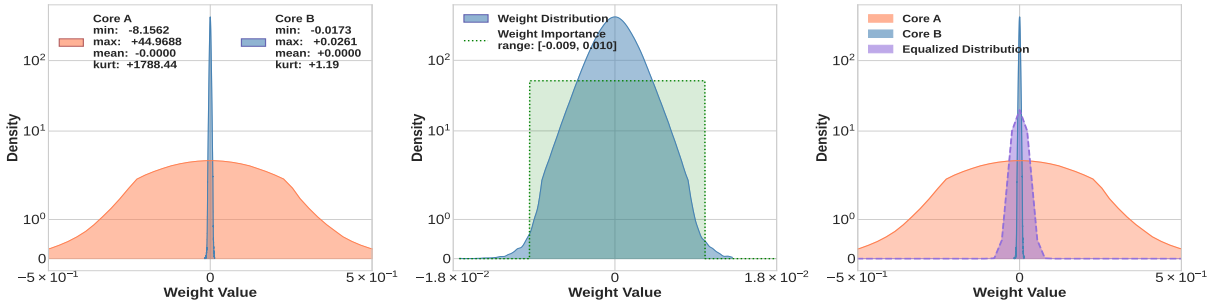


Figure 2: (a) Distribution of Core-A,B's weight parameters of different layers (b) Weights considered important on basis of estimated probability density (c) Equalized core distribution (purple) expected post core-equalization. The layer `model.layers.22.self_attn.o_proj` of Qwen3-8B is considered for these visualizations.

the ranges of A' and B' be $[A'_{\min}, A'_{\max}]$ and $[B'_{\min}, B'_{\max}]$, respectively. Since $A' = cA$ and $B' = \frac{B}{c}$, matching the corresponding ranges yields

$$c = \sqrt{\frac{B_{\max} - B_{\min}}{A_{\max} - A_{\min}}}. \tag{10}$$

This equalization strategy improves distribution alignment and stabilizes downstream adaptation, as quantifiably explored in Sec. 4.3.1.

However, we observe that several layers exhibit heavy tailed weight distributions, making full-range alignment sensitive to outliers. In such cases, aligning the dominant high-density regions is often more effective than matching the entire support. Motivated by this observation, we introduce a quantile-constrained variant that limits the influence of extreme tails while preserving the majority of important weights.

Strategy-ii: Quantile-Based Equalization To mitigate the influence of distributional tails, we redefine the effective support of each core using quantile-based ranges rather than the full minimum-to-maximum extent, yielding a more faithful representation of the core's importance distribution (operational weight-range). For core A', the effective range is thus defined as

$$A'_{\min} = F^{-1}_{A'}(\gamma_A) \tag{11}$$

$$A'_{\max} = F^{-1}_{A'}(1 - \gamma_A) \tag{12}$$

where $\gamma_A = p_x \cdot \tanh(\beta \kappa_A)$, ; p_x is the quantile level hyperparameter, κ_A denotes the kurtosis of $f_A(\cdot)$, β is a hyperparameter, and $F^{-1}_{A'}(\cdot)$ denotes the inverse Cumulative Distribution Function(CDF) of A'.

Note that scaling by c preserves kurtosis and scales $p-$quantiles proportionally, i.e., $F^{-1}_{A'}(p) = cF^{-1}_{A}(p)$. Using this property, the effective range of A' can be written as

$$\Delta'_A = c\left(F^{-1}_A(1-\gamma_A) - F^{-1}_A(\gamma_A)\right) \tag{13}$$

and similarly for Δ'_B. Matching these effective ranges yields

$$c = \sqrt{\frac{\Delta_B}{\Delta_A}}. \tag{14}$$

The $\tanh(\cdot)$ function is used to smoothly increase tail suppression with increasing kurtosis while remaining bounded within $[0, 1]$. This allows heavily tailed distributions to ignore outliers during alignment without discarding a significant fraction of important weights. As shown in Figure 2(c), this produces substantially better distribution alignment compared to full-range equalization.

Layer-wise Application - Note that, in both strategies, we compute the optimal scaling factor c^*_l independently for each layer l, accounting for the layer-specific weight distributions. This localized approach preserves the representational capacity of each layer while ensuring within-layer distribution alignment. The per-layer scaling factors are computed once during model initialization and applied deterministically, introducing no additional computational overhead during training or inference.
We use quantile limit $p_x = 5 \times 10^{-3}$ (atmost 0.5% of outlier weights can be ignored on either of the sides) and the kurtosis scaling factor $\beta = 10^{-3}$. However, we noted method stability under moderate variations of these hyperparameters (Table 6 of Appendix A.4). Following compression, we perform distillation powered short-horizon logit alignment using 5k effective calibration samples [3] from the FineWeb Dataset(Penedo et al., 2024).

[3] Due to sequence chunking in the data loader, a larger raw sample count is maintained to guarantee an effective size of 5,000 calibration samples.

As formulated in Section 3.3.1, core equalization operates as a weight-preserving transformation on the reconstructed tensors, meaning it induces no immediate change in standalone compressed model perplexity or base accuracy. However, its primary value lies in conditioning the representation space: incorporating core equalization significantly boosts downstream transferability and sharpens short-horizon logit alignment (Table 3). To contextualize these gains against existing literature, Table 4 benchmarks our approach against state-of-the-art (SOTA) compression methods, demonstrating superior logit-matching fidelity across evaluation horizons.

4 Results

In this section, we perform comprehensive studies to demonstrate the reliability, robustness, and generalizability of our proposed approach. We primarily evaluate our setup across various compression percentages on the Llama (Touvron et al., 2023a,b) and Qwen (Yang et al., 2025a) model families, establishing generalizability and robustness. We also provide detailed analysis of the effect of various chosen hyperparameters and design choices. We evaluate language modeling quality using perplexity on WikiText-2 (Merity et al., 2016) and C4 (Raffel et al., 2020). For Zero-shot task accuracy we use PIQA (Bisk et al., 2020), WinoGrande (Sakaguchi et al., 2021), OpenBookQA (Mihaylov et al., 2018), Arc-Easy, Arc-Challenge (Clark et al., 2018) and HellaSwag (Zellers et al., 2019) datasets.
Compression is applied to the linear projection matrices in both attention and MLP blocks, using truncated singular value decomposition. Unless otherwise stated, weights are stored in FP16, and no task specific training is performed after compression. All methods use identical evaluation settings.

4.1 Performance | Compression Methods

We compare GRACE against a diverse set of state-of-the-art compression approaches in Table 1, including learnable decomposition methods (e.g., ARA), manual decomposition techniques (e.g., SVD-LLM), and widely used pruning-based baselines. The comparison spans multiple compression regimes to evaluate model fidelity, robustness and performance preservation under varying levels of parameter reduction. For all comparative models, the results are as reported in the source works.

Table 1 compares Qwen3-8B performance across different compression ratios. Compared against ARA - the strongest adaptive-rank baseline, GRACE-SVD improves perplexity **by up to 25% under comparable compression budgets**, with gains becoming more pronounced at higher compression ratios. Noticeably, the language model capabilities of Qwen3-8B compressed by 60% (ratio 0.4) using GRACE-SVD is better than 0.6 ratio of existing baselines.

Beyond perplexity improvements, GRACE-SVD consistently achieves stronger zero-shot performance across reasoning and commonsense benchmarks. At 0.6 ratio, GRACE improves average zero-shot accuracy by approximately 25% over ARA, indicating better preservation of downstream reasoning capabilities under aggressive compression. Remarkably, GRACE-SVD preserves stable performance even at 0.4 parameter ratio, indicating effective retention of model quality under highly aggressive compression settings.

4.2 Performance | Model Families

To validate performance across different model families, we present our results on Llama3.1-8B in Table 2[4]. GRACE achieves substantial improvements in performance metrics, demonstrating robustness to variations in baseline architecture and training data, highlighting generalizability of the proposed pipeline. Compared to other decomposition methods, GRACE achieves a substantial improvement in perplexity retention. This trend continues under aggressive compression settings as well. With nearly 0.4 parameter ratio, GRACE manages to retain a significant portion of base model capabilities.
A key objective of core equalization is preserving the downstream adaptability of compressed representations. To validate this, detailed post-compression fine-tuning results are provided in Appendix A.6.

4.3 Ablation Study

4.3.1 Comparing different Core Equalization techniques

We compare different core equalization techniques explored within the GRACE pipeline. Specifically, we evaluate (i) Uniform scaling for all whitening

[4]ARA provides results only for Llama2 family models within Llama family, and hence could not be compared against Llama3.1 8B. We target expanding our setup to include this in the future. However, to further prove robustness, we present the results of compression on Qwen2.5-3B in Appendix A.7

Ratio	Category	Method	PPL ↓ Wiki2	C4	Acc. ↑ PIQA	Wino	OBQA	Arc-c	Arc-e	HellaS	Avg.
1.0	–	Dense	9.72	15.42	76.88	67.80	31.00	55.97	83.54	57.13	62.05
0.8	Pruning	DLP	15.67	42.60	68.28	63.85	28.00	42.32	68.14	43.60	52.37
	Non-Trainable Rank Selection	FARMS	14.44	38.72	69.64	66.06	27.80	41.47	69.53	44.27	53.13
	Trainable Rank Selection	Dobi-SVD	12.45	30.81	71.33	65.19	29.60	43.52	72.39	45.62	54.61
		ARA	10.75	20.05	73.78	67.01	29.80	52.39	80.43	49.34	58.79
		GRACE	**10.24**	**17.29**	**74.87**	**68.06**	**30.48**	**53.72**	**83.09**	**50.55**	**60.13**
0.6	Pruning	DLP	27.78	117.86	60.39	56.75	17.80	23.21	42.72	32.16	38.84
	Non-Trainable Rank Selection	FARMS	34.50	168.17	58.98	53.59	16.00	21.50	40.78	31.00	36.98
	Trainable Rank Selection	Dobi-SVD	18.65	64.08	63.60	59.59	24.40	29.52	55.13	35.90	44.69
		ARA	16.34	47.71	65.72	57.46	22.40	30.97	59.01	36.24	45.30
		GRACE	**12.17**	**19.12**	**71.33**	**65.42**	**28.84**	**48.96**	**79.00**	**47.84**	**56.90**
0.4	Trainable Rank Selection	**GRACE**	**13.91**	**21.55**	**70.05**	**63.20**	**27.22**	**44.74**	**76.16**	**44.13**	**54.25**

Table 1: Performance comparison on Qwen3-8B. We report PPL (Wiki2/C4) and Zero-shot accuracy (PIQA/Wino/OBQA/Arc-c/Arc-e/HellaS). Lower PPL is better; higher accuracy is better. Best results in **bold**.
DLP - (Chen et al., 2025), FARMS - (Hu et al., 2025)

Ratio	Method	PPL ↓ Wiki2	C4
1.0	Dense	6.14	9.47
0.8	FWSVD	4782	8195
	ASVD	17.55	28.41
	SVD-LLM	11.82	20.05
	SVD-LLM V2	8.01	11.72
	GRACE	**7.54**	**11.48**
0.6	**GRACE**	10.10	17.44
0.4	**GRACE**	12.32	19.25

Table 2: PPL comparison on LLaMA3.1-8B across compression ratios. Lower is better. Best results in **bold**.
FWSVD: (Hsu et al., 2022), ASVD: (Yuan and et al., 2024), SVD-LLM: (Wang et al., 2025c), SVD-LLM V2: (Wang et al., 2025b)

matrices using a globally selected scaling factor (ii) Norm based scaling (iii) Min-Max range based scaling as described in Section 3.3.2.i, and (iv) the proposed distribution-aware scaling approach based on quantile-based Wasserstein distance minimization (Section 3.3.2.ii).

The qualitative comparisons of these strategies is presented in Table 3, highlighting the impact of each on overall performance and model fidelity. Quantile-based CE reduces the drop in perplexity by $\approx 60\%$, providing a more stable model for downstream use. While we also benchmark against a standard Norm-Based CE [5] approach in Table 3, its empirical gains remain suboptimal. This limitation stems directly from the fact that norm-based methods only equalize aggregate energy across factors rather than local dynamic ranges; as a result, extreme-value outliers and core-wise range disparities remain largely unmitigated compared to our range- and quantile-aware formulations. We provide a concrete toy counterexample in Appendix A.5 mathematically demonstrating how identical matrix norms can mask dynamic range disparities.

Additionally, Table 6 in Appendix A.4 presents a study on the key hyperparameters used in the strategy formulation, analyzing the sensitivity of the method to different distribution truncation and scaling configurations.

Method	**Wiki2 ↓**	**PTB ↓**	**Wiki2 Drop from Base ↓**
Non Core Equalized	13.458	46.262	38.46%
Uniform Scaling CE	11.962	23.077	23.07%
Norm CE	11.772	23.077	21.11%
Min-Max Range CE	11.392	22.965	17.20%
Quantile-Based CE	**10.841**	**21.659**	**11.53%**

Table 3: Ablation study on different core equalization methods (refer Section 3.3.2) Qwen3-8B 0.7 Parameter Ratio compression. Base Wiki2 PPL = 9.72.

4.3.2 Direct VS Progressive Decomposition

We compare direct against a progressive decomposition strategy with intermediate core equalization after each stage. While direct decomposition applies the target reduction in a single step, progressive decomposition incrementally constrains the model, allowing periodic redistribution and recalibration of decomposed representations. Results in Fig 3 show that progressive decomposition consistently outperforms direct decomposition, with gains becoming more pronounced at higher compression ratios. At aggressive compression set-

[5] Here the scaling factor would be $c_{\text{norm}} = \sqrt{\frac{\|\mathbf{B}\|_F}{\|\mathbf{A}\|_F}}$

tings, progressive methods improve performance by up to 5% over direct methods, while maintaining lower perplexity and stronger downstream accuracy. These findings suggest that intermediate equalization mitigates the compounded degradation introduced by direct high-ratio compression.

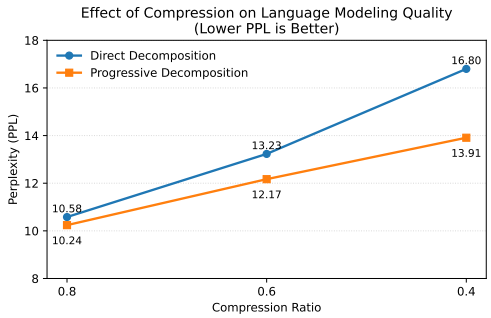


Figure 3: Wiki2 Perplexity comparison for direct decomposition versus progresive decomposition across Qwen-3 8B model. Progressive decomposition reaches the target ratio through intermediate compression ratios before the final step. We use the setup of 0.1->0.2... for progressive decomposition

4.3.3 Impact of logit alignment

As discussed in Section 3.3, we employ a short-horizon logit alignment stage to efficiently recover the initial performance drop inherent to any decomposition method. To isolate the efficacy of our guided rank-allocation and equalization methods from the effects of this post-compression calibration, we establish a controlled baseline comparison. We applied an identical logit alignment procedure to Dobi-SVD and ARA methodologies (using our own implementation). Table 4 details the results for the Qwen3-8B model at a 0.6 compression ratio. Under this standardized adaptation setup, GRACE-SVD continues to substantially outperform the baselines.

5 Conclusion and Future Work

We present **GRACE-SVD**, a trainable low-rank compression framework combining guided rank allocation, progressive decomposition, & core equalization to improve stability under aggressive compression. Across multiple model families and compression settings, GRACE consistently outperforms existing decomposition methods, maintaining lower perplexity & stronger zero-shot performance even at 0.4 parameter ratio, where competing methods often exhibit substantial degradation. Ablation studies further highlight the importance of trainable decomposition and distribution-aware equalization in preserving compressed model fidelity and adaptability.

Future work includes extending GRACE to Mixture-of-Experts and multimodal architectures, as well as developing unified end-to-end optimization strategies to investigate how rank-allocation strategies can be tailored to preserve the specific structural components necessary for multi-step logical reasoning, such as Chain-of-Thought (CoT) generation and mathematical problem-solving, which are often the most sensitive to matrix factorization. Furthermore, evaluating the interplay between matrix decomposition and advanced post-training alignment methods presents a vital area for ongoing research.

Method	Dobi-SVD	ARA	GRACE-SVD (Ours)
Wiki2 PPL ↓	16.41	13.34	**12.17**

Table 4: Ablation study on impact of logit alignment step on closest SOTA baselines

6 Limitations

GRACE-SVD is a trainable compression framework and therefore introduces additional optimization overhead compared to training-free decomposition methods. This computational cost further increases under progressive compression settings, where repeated decomposition and intermediate equalization improve robustness at the expense of compression time.

Additionally, the optimization objectives used in this work are primarily centered around next-token prediction. While empirical results show preservation of downstream performance, the framework does not explicitly optimize for retention of reasoning ability, factual knowledge, or instruction-following behavior under compression. We aim to solve this in future works. Finally, experiments in this work are limited to dense decoder-only language models. Extending GRACE-SVD to multimodal architectures, encoder-decoder models, and MoE systems remains an important direction for future work.

References

Aastik, Meghana Topu, Chinmay Kulkarni, and Pragya Paramita Sahu. 2025. NormAL LoRA: What

During the preparation of this work, the authors used a large language model (LLM) for assistance with paper structuring and grammar correction.

is the perfect size? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19716–19731, Suzhou, China. Association for Computational Linguistics.

Saleh Ashkboos, Maximilian Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. 2024. Slicegpt: Compress large language models by deleting rows and columns. In *International Conference on Learning Representations*, volume 2024, pages 11682–11701.

Cristian-Paul Bara, Qing Ping, Abhinav Mathur, Govind Thattai, Rohith MV, and Gaurav S. Sukhatme. 2022. Privacy preserving visual question answering. *Preprint*, arXiv:2202.07712.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. In *AAAI*.

Yuli Chen, Bo Cheng, Jiale Han, Yingying Zhang, Yingting Li, and Shuhao Zhang. 2025. DLP: Dynamic layerwise pruning in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 7934–7956. PMLR.

Cheng-Han Chiang and Hung-yi Lee. 2023. A closer look into using large language models for automatic evaluation. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 8928–8942.

Peter Clark, Isaac Cowhey, Oren Etzioni, and 1 others. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Elias Frantar and Dan Alistarh. 2023. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International conference on machine learning*, pages 10323–10337. PMLR.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. 2022. Language model compression with weighted low-rank factorization. *Preprint*, arXiv:2207.00112.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Yuanzhe Hu, Kinshuk Goel, Vlad Killiakov, and Yaoqing Yang. 2025. Eigenspectrum analysis of neural networks without aspect ratio bias. *arXiv preprint arXiv:2506.06280*.

Ranjeet Ranjan Jha, Andra Siva Sai Teja, Venkatesh Wadawadagi, and Ravindra Babu Tallamraju. 2024. Enhancing anomaly detection in noisy images: Unleashing the power of attention-aware pde constraint feature denoiser module. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9.

Xiaoqi Jiao and et al. 2020. Tinybert: Distilling bert for natural language understanding. *Findings of EMNLP*.

Xun Liang, Hanyu Wang, Huayi Lai, Simin Niu, Shichao Song, Jiawei Yang, Jihao Zhao, Feiyu Xiong, Bo Tang, and Zhiyu Li. 2025. Seap: Training-free sparse expert activation pruning unlock the brainpower of large language models. *Preprint*, arXiv:2503.07605.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of machine learning and systems*, 6:87–100.

Zichang Liu, Jue Wang, Tri Dao, Tianyi Zhou, Binhang Yuan, Zhao Song, Anshumali Shrivastava, Ce Zhang, Yuandong Tian, Christopher Re, and 1 others. 2023. Deja vu: Contextual sparsity for efficient llms at inference time. In *International Conference on Machine Learning*, pages 22137–22176. PMLR.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.

Thinh On, Hengzhi Pei, Leonard Lausen, and George Karypis. 2026. Knowledge distillation for large language models through residual learning. In *The Fourteenth International Conference on Learning Representations*.

Victor M. Panaretos and Yoav Zemel. 2019. Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, 6(1):405–431.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, and 1 others. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849.

Ruoling Qi, Yirui Liu, Xuaner Wu, Xiangyu Wang, Ming Li, Chen Chen, Jian Chen, Yin Chen, and Qizhen Weng. 2026. Swift-svd: Theoretical optimality meets practical efficiency in low-rank llm compression. *arXiv preprint arXiv:2604.01609*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Pragya Paramita Sahu, Abhishek Raut, Jagdish Singh Samant, Mahesh Gorijala, Vignesh Lakshminarayanan, and Pinaki Bhaskar. 2024. Pop-vqa – privacy preserving, on-device, personalized visual question answering. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8455–8464.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Mingjie Sun, Zhuang Liu, Anna Bair, and Zico Kolter. 2024. A simple and effective pruning approach for large language models. In *International Conference on Learning Representations*, volume 2024, pages 4942–4964.

Shicheng Tan, Weng Lam Tam, Yuanchun Wang, Wenwen Gong, Shu Zhao, Peng Zhang, and Jie Tang. 2023. Gkd: A general knowledge distillation framework for large-scale pre-trained language model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 134–148.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Qinsi Wang, Jinghan Ke, Masayoshi Tomizuka, Kurt Keutzer, and Chenfeng Xu. 2025a. Dobi-svd: Differentiable svd for llm compression and some new perspectives. In *International Conference on Learning Representations*, volume 2025, pages 12561–12590.

Wei Wang, Zheng Dang, Yinlin Hu, Pascal Fua, and Mathieu Salzmann. 2019. Backpropagation-friendly eigendecomposition. *Advances in Neural Information Processing Systems*, 32.

Xin Wang, Samiul Alam, Zhongwei Wan, Hui Shen, and Mi Zhang. 2025b. Svd-llm v2: Optimizing singular value truncation for large language model compression. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4287–4296.

Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. 2025c. SVD-LLM: Truncation-aware singular value decomposition for large language model compression. In *The Thirteenth International Conference on Learning Representations*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. 2024. Sheared llama: Accelerating language model pre-training via structured pruning. In *International Conference on Learning Representations*, volume 2024, pages 5385–5409.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. Smoothquant: Accurate and efficient post-training quantization for large language models. In *International conference on machine learning*, pages 38087–38099. PMLR.

Yuzhuang Xu, Xu Han, Zonghan Yang, Shuo Wang, Qingfu Zhu, Zhiyuan Liu, Weidong Liu, and Wanxiang Che. 2024. Onebit: Towards extremely low-bit large language models. *Advances in Neural Information Processing Systems*, 37:66357–66382.

Lin Xv, Jingsheng Gao, Xian Gao, Ting Liu, and yuzhuo fu. 2026. ARA: Adaptive rank allocation for efficient large language model SVD compression.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yifan Yang, Kai Zhen, Bhavana Ganesh, Aram Galstyan, Goeric Huybrechts, Markus Müller, Jonas M Kübler, Rupak Vignesh Swaminathan, Athanasios Mouchtaris, Sravan Babu Bodapati, and 1 others. 2025b. Wanda++: Pruning large language models via regional gradients. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4321–4333.

Zhihang Yuan and et al. 2024. Asvd: Activation-aware singular value decomposition for compressing llms. *arXiv preprint arXiv:2401.14124*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4791–4800.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*.

A Appendix

A.1 Layer-wise Rank Allocation Behaviour

The effectiveness of low-rank decomposition is highly dependent on both the layers selected for compression and the degree of compression applied to each layer. Fixed rank assignment strategies assume uniform sensitivity across layers, often leading to suboptimal reconstruction and downstream degradation. GRACE-SVD employs a trainable rank-allocation strategy guided by a combination of language modeling, compression, and guidance objectives. Rather than assigning absolute ranks, the optimization process learns layer-specific compression preferences under a target budget, allowing allocation to adapt to reconstruction sensitivity and downstream performance requirements. Figure-4 visualizes the resulting learned compression patterns across layers, when targeting a parameter ratio of 0.7. We observe substantial variation in retained ranks, with compression preferences differing even among neighboring layers. While certain middle-layer regions tend to receive stronger compression, the allocation remains non-uniform, suggesting that layer importance under decomposition cannot be captured through fixed heuristics alone. These observations reinforce the importance of Guided Rank Allocation and indicate that the combined optimization objective enables GRACE-SVD to identify more effective decomposition strategies under aggressive compression settings.

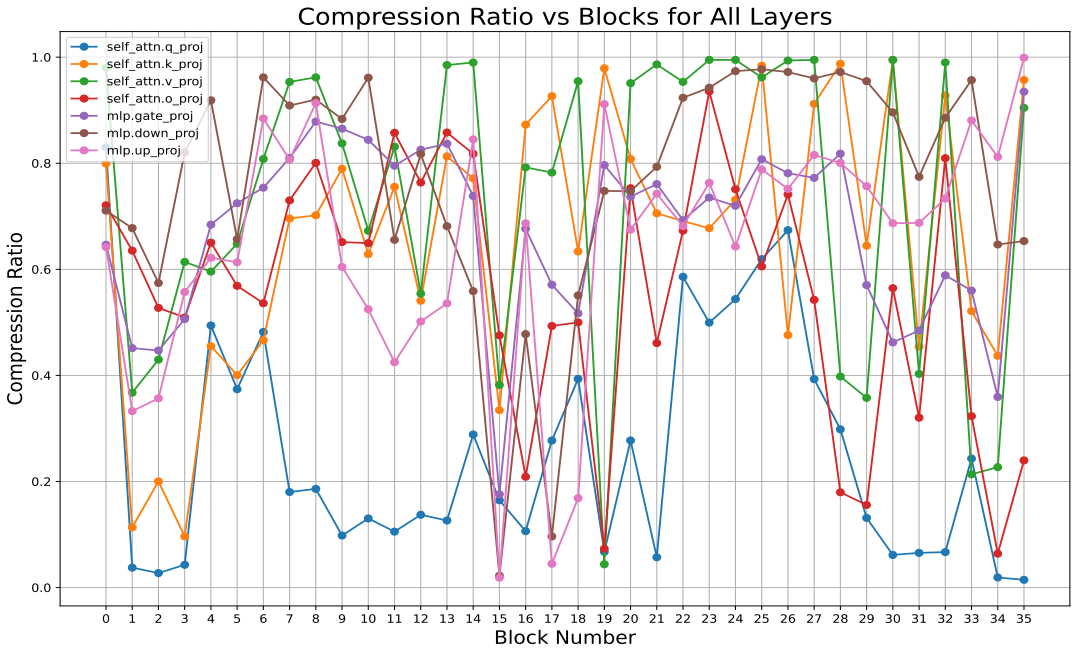


Figure 4: Variation in layer-wise compression to attain an effection target compression of 30% with QWEN3-8B

A.2 Training Hyperparameters and Optimization Dynamics

The guided rank-allocation pipeline is optimized using a weighted combination of language modeling, compression, and guidance objectives, balancing reconstruction quality with aggressive compression under a target budget. All the three components of the loss i.e. L_{ce}, L_{comp} and L_g can be understood from the following equation:

$$L = \lambda_{ce} \underbrace{\text{CE}\left(f\left(\boldsymbol{x}, \{\boldsymbol{\alpha}_i\}_{i=1}^{d}\right), \boldsymbol{y}\right)}_{L_{ce}} + \lambda_{comp} \underbrace{\left(R - R_{\text{target}}\right)}_{L_{comp}} + \lambda_g L_g \tag{15}$$

where, $\{\alpha_i\}_{i=1}^{d}$ represents the probability simplex as discussed in Section 3.2 , f is the LLM function, and (x, y) are the input text and the corresponding label. R represents the compression ratio as represented

in the Equation 5. Furthermore, the Guidance Loss L_g is calculated using a value, I_R representing the information retained by the model at the current compression ratio R. Thus, L_g is defined by:

$$L_g = \begin{cases} 0 & \text{if } I_R > (1 - R) \\ R & \text{if } I_R \leq (1 - R) \end{cases}$$

Hyperparameters were selected empirically to ensure stable optimization while preserving downstream performance. Table 5 summarizes the primary training settings used across experiments. We use Fully Sharded Data Parallel (FSDP) training to shard model parameters across devices, significantly reducing GPU memory requirements during optimization. Additionally, SVD is precomputed and stored along with whitening matrices such that loading these matrices is avoided repeatedly, thus reducing the number of Cholesky computations, improving runtime efficiency while reducing GPU memory overhead. All trainings are run on $4XA6000$ Nvidia GPUs.

Figure 5 illustrates the evolution of total loss L during training. We observe smooth convergence across competing objectives, indicating stable optimization of the rank-allocation process.

Table 5: Hyperparameters used for guided rank allocation training.

LR	λ_{ce}	λ_g	λ_{comp}	$\lvert\alpha\rvert$	Optimizer	Batch Size	Dataset
0.075	0.1	1.0	1.0	100	AdamW	32	WikiText

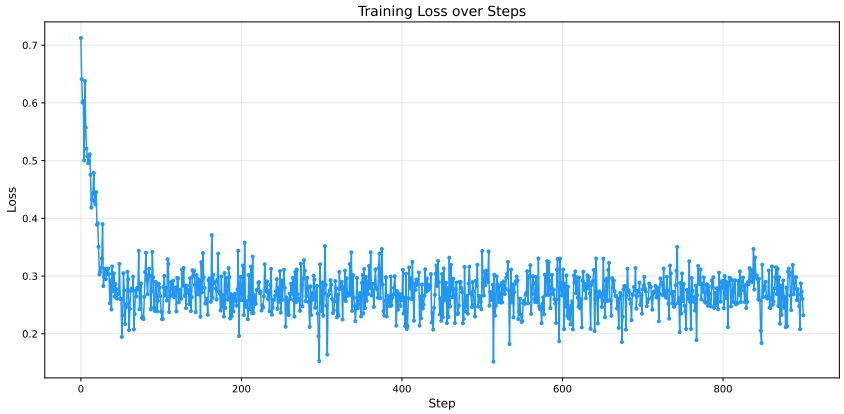


Figure 5: Loss curve for total loss L in Equation 15

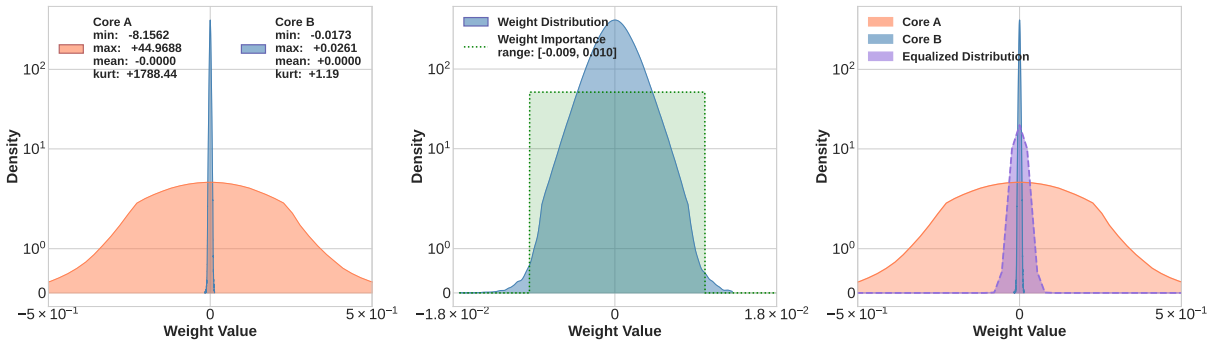


Figure 6: (a) Distribution of Core-A,B's weight parameters of different layers (b) Weights considered important on basis of estimated probability density (c) Equalized core distribution (purple) expected post core-equalization. The layer `model.layers.22.self_attn.o_proj` of Qwen3-8B is considered for these visualizations.

A.3 Core Distribution Analysis

In low-rank decomposed weights, core tensors corresponding to dominant singular values capture the most critical information. However, their parameter distributions vary significantly across layers and cores, causing imbalanced compression. Figure 6(a) shows this heterogeneity across Core-A and Core-B. We identify disproportionately important parameters via their estimated probability density (Figure 6(b)), and our core equalization redistributes them into a balanced representation (Figure 6(c), purple), preventing any single core from dominating the compressed model.

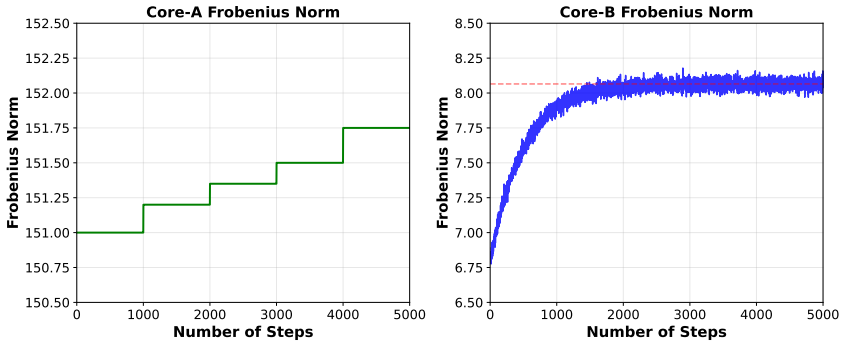


Figure 7: Variation in Frobenius Norm of decomposed cores in `model.layers.5.mlp.up_proj` layer of Qwen3-8B during fine-tuning on summarization dataset. Observe that Core-A is not learning when compared to Core-B. Core-A has just 0.5% change in norm - which is just a minor noise.

A.4 Core Equalization Hyperparameter Sensitivity

Core equalization introduces two hyperparameters: the probability threshold p_x (core fraction) and the smoothing strength β. We sweep $p_x \in \{10^{-4}, 5\times10^{-3}, 10^{-3}, 10^{-2}\}$ and $\beta \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$ on Qwen3-8B at 0.7 parameter ratio (Table 6). Too-small p_x (10^{-4}) under-equalizes; too-large (10^{-2}) over-smooths. The optimum at $p_x = 5\times10^{-3}$, $\beta = 10^{-3}$ yields the best Wiki2 PPL of 10.841.

p_x	β			
	10^{-1}	10^{-2}	10^{-3}	10^{-4}
10^{-4}	11.207	11.201	11.193	11.186
5×10^{-3}	10.986	10.968	**10.841**	10.849
10^{-3}	10.985	11.111	11.110	11.110
10^{-2}	11.228	11.193	11.192	11.194

Table 6: Sensitivity analysis of hyperparameters p_x and β on WikiText-2 perplexity of Qwen3-8B 0.7 Parameter Ratio compressed model, which were used in Core Equalization Method 3.3.2.

A.5 Toy Counterexample: Failure of Norm-Based Equalization

To illustrate the fundamental limitation of norm-based equalization, consider two decomposed factor matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{2\times2}$:

$$\mathbf{A} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} \sqrt{2} & 0 \\ 0 & -\sqrt{2} \end{bmatrix}. \tag{16}$$

Norm-Blindness of Standard Balancing: The Frobenius norms of both matrices are identical:

$$\|\mathbf{A}\|_F = \sqrt{1^2 + (-1)^2 + (-1)^2 + 1^2} = \sqrt{4} = 2, \tag{17}$$

$$\|\mathbf{B}\|_F = \sqrt{(\sqrt{2})^2 + 0^2 + 0^2 + (-\sqrt{2})^2} = \sqrt{4} = 2. \tag{18}$$

Under a norm-based equalization method, the scaling factor c is determined by the norm ratio:

$$c_{\text{norm}} = \sqrt{\frac{\|\mathbf{B}\|_F}{\|\mathbf{A}\|_F}} = \sqrt{\frac{2}{2}} = 1. \tag{19}$$

Because $c_{\text{norm}} = 1$, the transformations degenerate to the identity ($\mathbf{A}' = c\mathbf{A} = \mathbf{A}$ and $\mathbf{B}' = \mathbf{B}/c = \mathbf{B}$). Consequently, norm balancing performs no equalization, leaving the substantial dynamic range mismatch between $\mathbf{A} \in [-1, 1]$ and $\mathbf{B} \in [-\sqrt{2}, \sqrt{2}]$ completely unresolved.

Resolution via Range-Aware Equalization: Our proposed methods (Section 3.3.2 determines the scaling factor c using the dynamic bounds of the matrix entries:

$$c = \sqrt{\frac{B_{\max} - B_{\min}}{A_{\max} - A_{\min}}} = \sqrt{\frac{2 \cdot \sqrt{2}}{2 \cdot 1}} = \sqrt{\sqrt{2}} = 2^{1/4}.$$

Applying this scaling yields the transformed cores $\mathbf{A}' = c\mathbf{A}$ and $\mathbf{B}' = \frac{1}{c}\mathbf{B}$:

$$\mathbf{A}' = \begin{bmatrix} 2^{1/4} & -2^{1/4} \\ -2^{1/4} & 2^{1/4} \end{bmatrix}, \qquad \mathbf{B}' = \begin{bmatrix} 2^{1/4} & 0 \\ 0 & -2^{1/4} \end{bmatrix}.$$

Both $\mathbf{A}'$ and $\mathbf{B}'$ now share the exact balanced operational range $[-\sqrt{\sqrt{2}}, \sqrt{\sqrt{2}}]$ while strictly preserving the underlying reconstructed product $\mathbf{A}'\mathbf{B}' = (c\mathbf{A})(\frac{1}{c}\mathbf{B}) = \mathbf{AB}$.

Empirical Impact: This theoretical distinction directly reflects in downstream language modeling performance. On QWEN3-8B under a 0.7 parameter compression ratio, our quantile-based core equalization achieves a WikiText-2 perplexity of **10.841**, significantly outperforming norm-based equalization (**11.772**) by a verified 7.9% relative reduction and 10% overall reduction with respect to baseline in perplexity degradation.

A.6 Fine-Tuning Stability of Compressed Models

To further evaluate the quality of compressed representations, we study the post-compression supervised fine-tuning behavior. While reconstruction metrics & perplexity capture compression fidelity, downstream adaptation provides valuable insight into the trainability and representational robustness of compressed models. In particular, we analyze whether core equalization improves optimization stability during downstream adaptation.

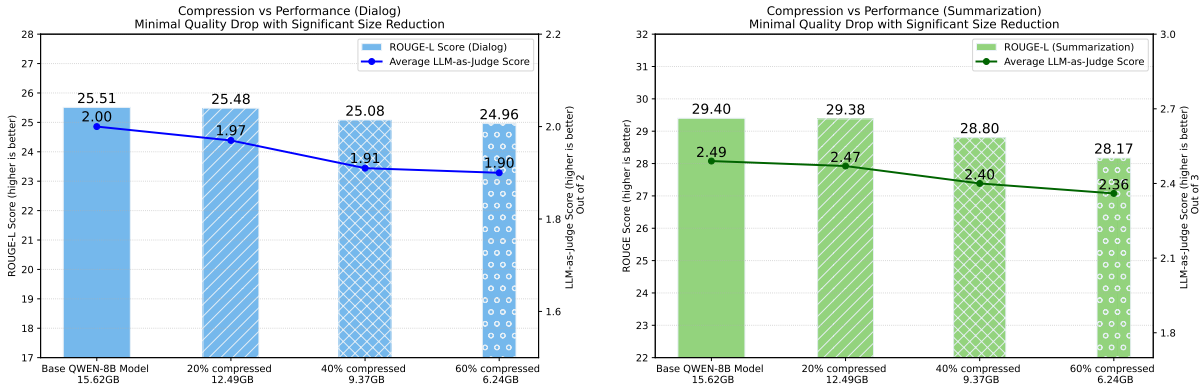
We fine-tune compressed models on summarization and dialogue generation tasks using a random subset of $200K$ samples from Flan-Dialog (Wei et al., 2021) and summarization datasets for one epoch. We used Adam Optimizer with a batch-size of 32 and a learning rate of $1e-5$. Performance is evaluated using ROUGE-L and LLM-as-a-judge scores (Fig.8). Across all compression settings, GRACE-SVD rapidly adapts to new tasks, preserving strong downstream performance even when initialized from compressed non-instruction models. These observations suggest that GRACE-SVD does not merely preserve reconstruction quality, but also maintains sufficient representational capacity for efficient downstream optimization through SFT methods; making it highly amenable to practical deployment in edge-applications.

A.7 Generalization to Smaller Dense Models: Qwen3-3B

To evaluate the robustness of GRACE-SVD across model scales, we apply the proposed pipeline to Qwen3-3B, a smaller dense language model. Under automatic rank selection with moderate compression, GRACE-SVD maintains competitive perplexity. These results indicate that the proposed compression framework remains effective even for lower-capacity dense models, preserving strong compression-performance tradeoffs. We further study whether compressed smaller models retain the ability to adapt to downstream tasks, using the previously described setup. Despite compression, GRACE-SVD preserves strong downstream adaptability, with compressed models achieving performance close to fine-tuned dense baselines across automatic and judge-based evaluations.

A.8 Qualitative Summarization and Dialog results

Tables 8 and 9 present qualitative examples for dialog and summarization tasks, respectively. Both the fine-tuned versions of Dense Qwen3-8B model and the 0.8 parameter ratio GRACE-SVD model produce



(a) Performance on dialog task

(b) Performance on summarization task

Figure 8: Task-specific comparison after finetuning of GRACE-Qwen models (VRAM in axis labels). Prompts used for LLM-as-judge are as defined in (Chiang and Lee, 2023) . We use GPT-OSS-120B for evaluation

Ratio	PPL (WikiText)	Summarization		Dialog	
		ROUGE-L	LLM-as-Judge	ROUGE-L	LLM-as-Judge
1.0	6.89	27.45	2.28	26.38	1.90
0.85	7.94	26.41	2.07	24.29	1.82

Table 7: Results on Qwen3-3B. PPL is computed post GRACE. Task specific scores are computed post fine-tuning as detailed in Main paper

fluent, factually grounded outputs. The compressed model occasionally omits secondary details or shifts factual emphasis, but does not introduce hallucinations—reflecting a natural salience variation from the information bottleneck rather than a quality degradation, consistent with our quantitative results.

Dialog Context	**Dense Model**	0.8 **Ratio (GRACE-SVD)**
Anonymous 1) What are the Panismahas? Anonymous 2) Panismahas The Panismahas or Panimaha were a subgroup of the Pawnee. Anonymous 1) When did they settle in the US? Anonymous 2)	The Panismahas were part of the Pawnee migration to the Platte River valley in the 17th century.	They settled in the US in the 1870s.
A dialog between 2 people: [+] What are the main types of communities and social groups in Australia? [+] Madjandji The Madjandji, also known as the "Majañji," are an indigenous Australian people in the area south of Cairns in the state of Queensland. [+] What language did the group speak? [+]	They speak the Madjandji language, which is part of the Pama-Nyungan language family.	They speak the Madjandji language, a dialect of the Yidini language.
– Who is Grete Jalk? – Grete Juel Jalk (1920–2006) was a Danish furniture designer. – What are some of her notable achievements? – Did she achieve any other things?	She was one of the first women to become a member of the Council of Industrial Design in Denmark.	She was the first woman to be appointed head of the Royal Danish Cabinet of Fine Arts in 1972.
+Where does David Malpass live? +Malpass and his wife, Adele, live in New York City. +What is the most interesting thing about David Malpass? +Adele Malpass was appointed as the Chairwoman of the Manhattan Republican Party in January 2015.	The Manhattan Republican Party is the local party organization of the Republican Party in New York City.	The Manhattan Republican Party is the largest and most influential Republican Party organization in New York City.

Table 8: Qualitative comparison of dialog responses between the dense model and the GRACE-SVD model. Both models generate coherent and grounded responses, demonstrating that GRACE-SVD preserves generative quality even at significant compression ratios.

Article	Dense Model	0.8 **Ratio using GRACE-SVD**
Martina Hingis will play her first singles match since 2007 on Saturday when they faces Poland's Agnieszka Radwanksa in the Fed Cup for Switzerland. Hingis won all five of her Grand Slam titles by the age of 18 and had the world at her feet, before retiring at the age of 22 in February 2003 due to injuries…	Martina Hingis will play her first singles match since 2007 on Saturday. Hingis won all five of her Grand Slam titles by the age of 18. The 34-year-old was ranked No 1 in the world for 209 weeks. Hingis retired in 2003 due to injuries but came back in 2005. She was banned for two years in January 2008 after failing a doping test. Hingis returned again at Wimbledon 2013 and has been playing doubles.	Martina Hingis will play her first singles match since 2007 on Saturday. The 34-year-old will face Agnieszka Radwanksa in the Fed Cup for Switzerland. Hingis won all five of her Grand Slam titles by the age of 18. The 34-year-old was ranked No 1 in the world for 209 weeks in her prime.
A father whose 20-year-old daughter was found murdered in Iowa last year made a desperate plea from China for U.S. authorities to do more in tracking down her killer, six months after the girl's body was discovered…	Tong Shao, 20, went missing in September 2014 and her body was found three weeks later in her car trunk. Police say her boyfriend, Xiangnan Li, 23, is wanted for her murder. He is believed to have been the last person to see her alive. Li left Iowa less than two days after she went missing and flew to China. Her father has pleaded with U.S. authorities to do more to find her killer. Li is wanted for murder in the U.S. but there is no extradition treaty with China.	Tong Shao, 20, went missing in September 2014 and her body was found in her car in Iowa City. Police say her boyfriend, Xiangnan Li, 23, was the last person to see her alive. He left Iowa for China two days after she went missing and has not been heard from since. A warrant has been issued for his arrest in the murder of Tong Shao.
Sam Bennett has revealed the extent of the shocking injuries he sustained after the Bora-Argon 18 cyclist was involved in 50-rider pile-up in the final kilometre of the Scheldeprijs one day Classic…	Sam Bennett was involved in a 50-rider pile-up at the Scheldeprijs. The 24-year-old was taken to hospital after the crash in Belgium. Bennett suffered cuts to his back and shoulders after the pile-up. Team Sky's Elia Viviani also collided with a barrier and was taken to hospital for x-rays complaining of pain in his arm. Alexander Kristoff of Katusha won the race after avoiding the crash.	Sam Bennett was involved in a pile-up in the final kilometre of the race. The 24-year-old was taken to hospital after the crash in Belgium on Monday. Bennett suffered severe cuts to his back and shoulders in the crash. Team Sky's Bradley Wiggins and Geraint Thomas were both riding in the race but managed to avoid the crash.
Judge Mary Stacey admitted the victim had been scarred psychologically by the sexual assault but freed Josh Sweet after serving just two months on remand. A burglar who broke into a young teacher's home and sexually assaulted her as she lay in bed…	Joshua Sweet, 20, broke into the home of a 24-year-old teacher. He sexually assaulted her as she lay in bed, before fleeing. The victim was left petrified and traumatised by the ordeal. Judge Mary Stacey admitted the victim had been scarred psychologically. But she freed Sweet after he served just two months on remand. He was given a three-year community order, 150 hours unpaid work and must register as a sex offender for five years.	Joshua Sweet, 20, broke into the victim's home and sexually assaulted her. Judge Mary Stacey admitted the victim had been scarred psychologically. But Sweet was freed after serving just two months on remand. Victim Support charity worker said victims deserve justice.

Table 9: Qualitative comparison of summarization outputs between the dense model and the GRACE-SVD model. The GRACE-SVD model preserves the core factual content and coherence of the summaries, demonstrating the effectiveness of the proposed compression strategy. Article texts are truncated for brevity (indicated by …).