HiSVD: Principled Low-Rank Approximation of LLMs via Hierarchical Modeling of Information Capacity and Spectral Structure

Zhuo Chen[1,2], Minghao Li[2], Xiaoqian Ma[2], Siqi Fan[2], Xiusheng Huang[2], ZhangLiujie[2*], Weihang Chen[2*]

[1]Shanghai Jiao Tong University
[2]Xiaohongshu Inc.

Correspondence: zhangliujie@xiaohongshu.com, chenjinzhi@xiaohongshu.com

Abstract

Singular Value Decomposition (SVD) enables hardware-agnostic LLM compression via low-rank approximation, yet optimal rank allocation remains a bottleneck. Existing methods predominantly derive layer importance from performance-oriented proxies. Yet, these metrics fail to distinguish between representational importance and structural compressibility, consequently obscuring the fine-grained influence of spectral distribution shape. We demonstrate this disconnect through spectral analysis, revealing that layers with similar information capacity can exhibit markedly different singular value decay behaviors, corresponding to varying degrees of redundancy in the spectral tail. This imperfect coupling implies that allocation strategies driven solely by importance leave significant compression opportunities underexploited. To address this gap, we propose HiSVD, a hierarchical rank allocation framework with two stages: **(1) Capacity-Anchored Baseline Allocation**, which preserves representational stability by aligning rank budgets with information capacity; and **(2) Redundancy-Aware Refinement**, which modulates this baseline using tail redundancy to penalize structural excess. Experiments on LLMs demonstrate that HiSVD achieves superior compression efficiency, significantly outperforming state-of-the-art baselines by effectively exploiting this spectral heterogeneity.

1 Introduction

Pre-trained Large Language Models (LLMs) demonstrate remarkable capabilities but suffer from prohibitive memory and latency costs (Zhao et al., 2024; Brown et al., 2020; Patterson et al., 2021; Strubell et al., 2019). To mitigate these costs, model compression has become imperative, with techniques ranging from quantization (Dettmers et al., 2023; Shao et al., 2024) and pruning (Frantar and Alistarh, 2023; Sun et al., 2024) to knowledge distillation (Hsieh et al., 2023). Among these, Singular Value Decomposition (SVD)-based low-rank approximation (Golub et al., 1987) offers a unique hardware-agnostic advantage, directly translating parameter reduction into inference acceleration. While recent methods like FWSVD (Hsu et al., 2022), AdaSVD (Li et al., 2025), and DobiSVD (Wang et al., 2025) have improved SVD by optimizing truncation objectives (e.g., sensitivity, reconstruction error), they remain fundamentally constrained by the Rank Allocation bottleneck.

Specifically, existing methods predominantly rely on performance-oriented proxies (e.g., reconstruction error, sensitivity) to infer compressibility. However, these metrics often fail to distinguish between representational necessity and structural redundancy. This practice induces a conservative bias—layers exhibiting high sensitivity or large spectral energy are systematically protected with higher ranks—regardless of their actual structural properties. Here, Information Capacity establishes the fundamental representational demand of a layer, whereas Spectral Compressibility reflects the extent to which its weight matrix admits low-rank approximation. Crucially, spectral distribution shape is distinct from layer importance; instead, it characterizes the structural degrees of freedom that govern compressibility, even for layers with high information capacity. As a result, existing approaches often overlook fine-grained variations in spectral distribution shape—particularly tail redundancy—thereby obscuring potential compression opportunities that performance-driven signals alone fail to expose.

We investigate this disconnect in Figure 1. Figure 1(a) presents the singular value decay curves of two representative layers whose Effective Ranks are nearly identical. Despite this similarity, their spectral decay behaviors diverge substantially beyond the leading components, indicating that layers with similar information capacity can differ markedly in spectral tail redundancy. To assess whether this phenomenon is pervasive, we extend

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20659–20677

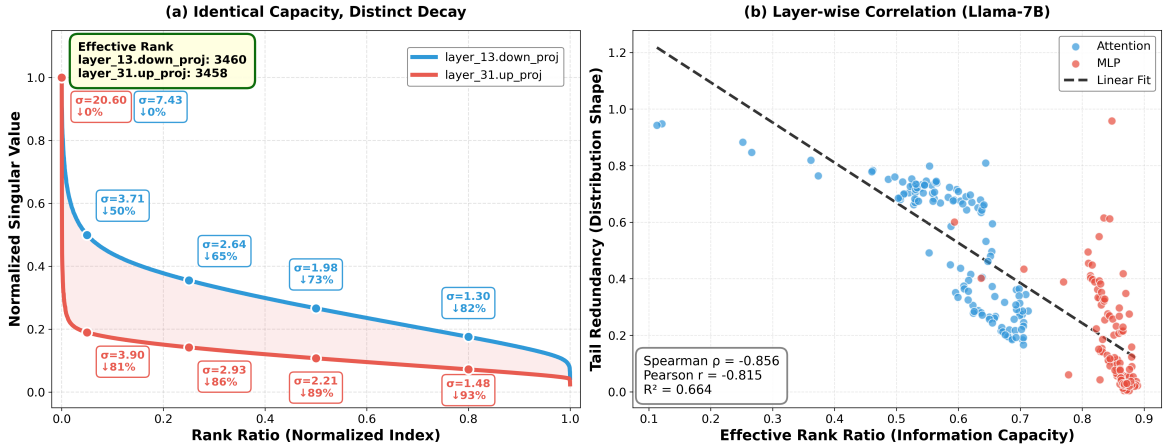

Figure 1: Imperfect Coupling between Information Capacity and Spectral Distribution Shape. (a) Two layers with nearly identical Effective Rank exhibit substantially different singular value decay patterns, revealing distinct structural redundancy despite comparable capacity. (b) Scatter plot across all LLaMA-7B linear layers shows strong global negative correlation (Spearman $\rho \approx -0.86$), yet moderate linear fit ($R^2 \approx 0.66$) indicates significant vertical dispersion: layers with similar information capacity can display widely varying degrees of tail redundancy. This motivates explicitly decoupling capacity from distribution shape. Comparative analysis is provided in Appendix C.

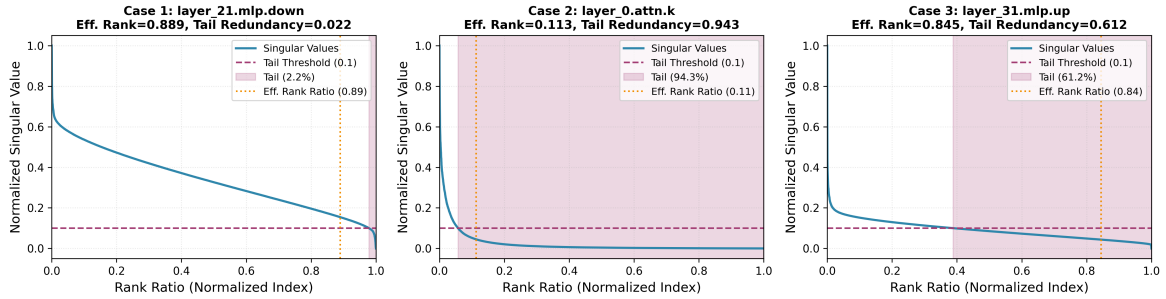

Figure 2: Representative Spectral Patterns Illustrating Structural Inflation. Case 1 corresponds to a hard-to-compress layer, while Case 2 is trivially compressible. Crucially, Case 3 reveals a structurally inflated layer: despite having an effective rank ratio comparable to Case 1, it exhibits pronounced tail redundancy similar to Case 2. This anomaly motivates the need to explicitly decouple information capacity from spectral distribution shape.

our analysis to all linear layers in the LLaMA family (Fig. 1(b)). At a macroscopic level, a strong global negative correlation (Spearman $\rho \approx -0.86$) indicates a pronounced monotonic trend: layers with higher information capacity tend to exhibit lighter spectral tails. This observation explains why existing compression strategies—guided by sensitivity, activation statistics, or reconstruction objectives—are often effective in a coarse, aggregate sense. However, a linear regression analysis ($R^2 \approx 0.66$) shows that a substantial portion of the variability in spectral distribution shape remains unexplained by capacity metrics alone. Importantly, this discrepancy stems not merely from nonlinearity but from pronounced vertical dispersion: layers with nearly identical information capacity frequently exhibit drastically different degrees of tail redundancy. While prior methods may incorporate multiple signals to estimate layer importance, they generally lack an explicit mechanism to disentangle information capacity from spectral distribution shape. Consequently, layers with comparable capacity but distinct structural redundancy are often treated similarly, leaving a significant compressibility gap underexploited.

Figure 2 crystallizes the practical implication of this gap. While low-capacity layers are naturally compressible (Case 2) and compact high-capacity layers require protection (Case 1), a central blind spot exists: Case 3. These layers possess high Information Capacity—appearing "important" to conventional metrics—yet exhibit heavy spectral tails similar to Case 2. Importance-driven heuristics typically over-allocate ranks to Case 3, missing the opportunity to prune its redundant degrees of freedom without compromising information integrity.

Driven by these insights, we propose HiSVD, a principled hierarchical rank allocation framework that explicitly disentangles Information Capacity from Spectral Distribution Shape. First, **Capacity-Anchored Baseline Allocation** establishes a foundational rank assignment derived from Effective Rank, capturing macroscopic representational requirements to ensure global stability. Second, **Redundancy-Aware Refinement** serves as a theoretically grounded modulation using fine-grained Tail Redundancy statistics; this step aligns rank perturbations with marginal spectral energy to rigorously minimize reconstruction error. By anchoring allocation on capacity and refining via structural redundancy, HiSVD systematically exploits latent compressibility that remains inaccessible to purely capacity-centric or ad-hoc heuristic approaches. Extensive experiments demonstrate that HiSVD consistently outperforms state-of-the-art baselines, validating the necessity of an interpretable, hierarchical decoupling of spectral shape from layer importance for efficient low-rank compression.

2 Related Works

2.1 SVD-based LLM Compression

Singular Value Decomposition has emerged as a promising post-training compression strategy for LLMs, offering hardware-agnostic acceleration that directly translates parameter reduction into inference speedup. Recent methods have significantly improved truncation quality by incorporating various performance-oriented optimization signals. FWSVD (Hsu et al., 2022) employs Fisher information weighting to guide parameter-level truncation decisions. ASVD (Yuan et al., 2023) exploits the observation that activation matrices exhibit stronger low-rank structure than weight matrices, decomposing features rather than weights for more accurate approximations. SVD-LLM (Wang et al., 2024) establishes explicit relationships between singular value magnitudes and compression error through data-dependent whitening transformations. While these methods demonstrate strong empirical performance, they share a common characteristic: they improve truncation through signals—reconstruction error, activation norms, Fisher information—that implicitly serve as proxies for compressibility, without explicitly modeling how spectral distribution shape affects structural compression capacity.

2.2 Rank Allocation and Layer Importance Estimation

Determining optimal compression budgets across heterogeneous layers represents a central challenge in structured compression. Traditional search-based strategies, including reinforcement learning (Schulman et al., 2017) and evolutionary optimization (Real et al., 2019), suffer from computational overhead that limits scalability. Modern differentiable frameworks mitigate this: ARS (Gao et al., 2024) introduces binary masking for differentiable rank optimization; AdaSVD (Li et al., 2025) assigns layer-specific ratios based on sensitivity metrics with adaptive compensation; DobiSVD (Wang et al., 2025) employs differentiable truncation to jointly optimize allocation and reconstruction through gradient-based search. However, these methods typically allocate ranks in a manner that correlates compression budgets with layer importance, implicitly assuming that layers deemed more critical should be protected with proportionally higher ranks. This assumption becomes insufficient when importance diverges from compressibility.

2.3 Spectral Analysis and Redundancy Metrics

Analyzing matrix redundancy is essential for understanding model compressibility. Classical metrics like Effective Rank (Roy and Vetterli, 2007) and Stable Rank (Vershynin, 2018; Cohen et al., 2016) quantify information content by aggregating singular values into scalar indices. While useful for estimating Information Capacity, these scalar summaries fail to capture the *shape* of the spectral distribution—specifically, the distinction between head-concentration and tail-heaviness. Although recent studies have linked spectral decay patterns to generalization (Martin and Mahoney, 2021) or fine-tuning stability (e.g., WeLORE (Jaiswal et al., 2025)), these insights have not been fully operationalized for compression. Existing frameworks either aggregate spectra into scalar importance proxies or optimize reconstruction without explicitly distinguishing capacity from distribution shape. In contrast, our work explicitly decouples spectral distribution shape from information capacity, utilizing tail redundancy as a distinct refinement signal to modulate capacity-anchored rank allocation, thereby transcending the constraints of traditional importance metrics.

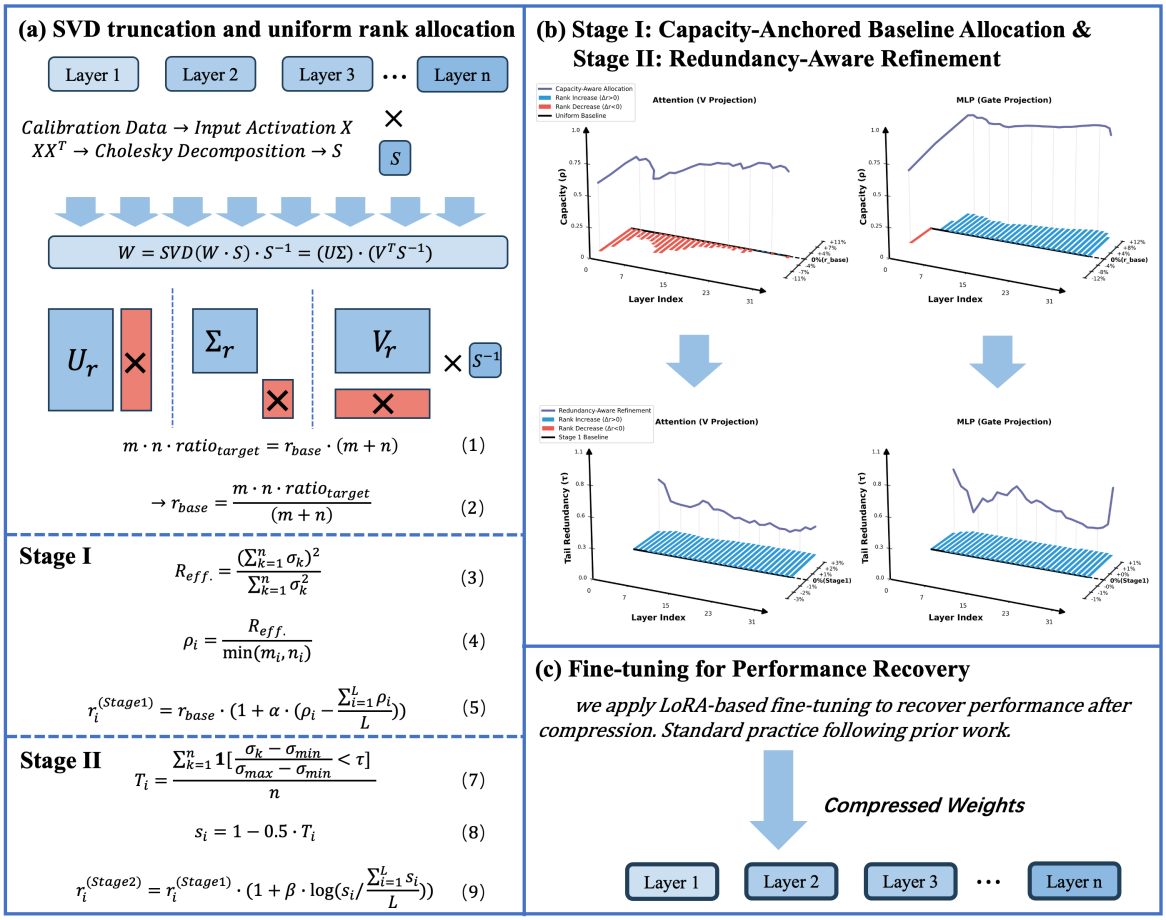

Figure 3: Overview of HiSVD. HiSVD performs hierarchical rank allocation in two stages: Stage I (Capacity-Anchored Baseline Allocation) assigns ranks according to effective rank ratio, capturing global differences in information capacity; Stage II (Redundancy-Aware Refinement) further modulates the baseline using tail redundancy to compress structurally inflated layers that deviate from this trend. Pipeline is shown in Algorithm E.

3 HiSVD

3.1 Motivation

Consider a pre-trained LLM with L linear layers, where each layer contains a weight matrix $W \in \mathbb{R}^{m \times n}$. SVD-based compression factorizes $W = U\Sigma V^\top$ and truncates the decomposition at rank r, reducing the number of parameters from mn to $r(m+n)$. Given a target compression ratio $ratio_{\text{target}}$, the rank allocation problem seeks to assign a rank r_i to each layer i such that the overall compressed model satisfies the budget constraint:

$$\sum_{i=1}^{L} r_i(m_i + n_i) = ratio_{\text{target}} \cdot \sum_{i=1}^{L} m_i n_i, \quad (1)$$

where m_i and n_i denote the dimensions of the weight matrix in layer i.

A straightforward baseline allocates ranks uniformly according to the global compression ratio. For a weight matrix $W \in \mathbb{R}^{m \times n}$, the baseline rank is given by:

$$r_{\text{base}} = \left\lfloor \frac{mn \cdot ratio_{\text{target}}}{m+n} \right\rfloor, \quad (2)$$

where $\lfloor \cdot \rfloor$ denotes the floor operator. While simple and budget-consistent, this uniform allocation treats all layers equally, ignoring their heterogeneous importance and spectral structure.

Recent methods improve upon uniform rank allocation by incorporating performance-oriented signals, such as sensitivity measures or reconstruction error, to guide rank assignment across layers. These approaches effectively differentiate layer importance, yet they largely operate through indirect proxies and do not explicitly account for how spectral distribution shape influences compressibility.

As revealed by the empirical observations in Section 1, this omission becomes non-negligible. Even among layers with comparable information capacity, their singular value decay profiles can differ substantially, exhibiting markedly different degrees

of tail redundancy. This discrepancy indicates that information capacity alone does not uniquely determine compressibility, and that spectral distribution shape introduces additional structural degrees of freedom that are not captured by importance-oriented proxies.

Motivated by this phenomenon, HiSVD formulates rank allocation as a two-stage hierarchical process. This design provides interpretability by physically separating the allocation logic: Stage I anchors allocation to the global capacity trend, while Stage II targets specific structural redundancies, allowing the compression strategy to be understood in terms of spectral properties rather than opaque optimization scores.

3.2 HiSVD: Hierarchical Rank Allocation

HiSVD performs rank allocation through a two-stage hierarchical procedure that separates information capacity estimation from structural redundancy modeling. The first stage establishes a capacity-anchored baseline allocation across layers, while the second stage refines this allocation by explicitly accounting for excess spectral redundancy reflected in the spectral tail.

Stage I: Capacity-Anchored Baseline Allocation

We begin by establishing a baseline allocation rooted in the relative information capacity of each layer. This foundational step is critical for preserving representational stability: it ensures that layers encoding high-dimensional intrinsic information (i.e., high Effective Rank) are allocated commensurate budgets prior to any redundancy reduction. By anchoring the rank distribution to information capacity, we prevent the subsequent refinement stage from inadvertently truncating essential spectral components in information-dense layers. Information capacity is quantified using the Effective Rank, which measures the intrinsic dimensionality of a weight matrix independent of its physical size. Given a weight matrix W with singular values $\{\sigma_k\}_{k=1}^{n}$, the effective rank is defined as

$$R_{\text{eff}} = \frac{\left(\sum_{k=1}^{n} \sigma_i\right)^2}{\sum_{k=1}^{n} \sigma_i^2}. \tag{3}$$

We utilize this participation-ratio-based effective rank as a lightweight proxy for capacity. We emphasize that this choice is a practical and interpretable proxy rather than a theoretically optimal notion of rank. a discussion on metric selection is provided in Appendix A.

To enable consistent comparison across layers of different shapes, we quantify the normalized information capacity using the effective rank ratio:

$$\rho_i = \frac{R_{\text{eff}}}{\min(m_i, n_i)}, \quad \rho_i \in (0, 1]. \tag{4}$$

Starting from a uniform baseline rank r_{base}, we reallocate ranks according to deviations in information capacity:

$$r_i^{(Stage1)} = r_{\text{base}} \left(1 + \alpha \cdot (\rho_i - \bar{\rho})\right), \tag{5}$$

where $\bar{\rho} = \frac{1}{L}\sum_{i=1}^{L} \rho_i$ denotes the mean normalized information capacity across layers, and α controls the strength of capacity-anchored allocation. This zero-mean formulation preserves the global compression budget while assigning relatively higher ranks to layers with greater information capacity and fewer ranks to layers with lower information capacity.

Stage II: Redundancy-Aware Refinement

The baseline allocation derived from capacity estimation treats all layers with similar effective ranks equally. However, this overlooks structural redundancy—the physical phenomenon where a layer possesses excess degrees of freedom despite high capacity. To address this, we introduce a refinement stage driven by tail redundancy, a quantitative metric we define to measure this structural excess. Inspired by the low-rank estimation strategy in WeLORE (Jaiswal et al., 2025), we operationalize tail redundancy by calculating the proportion of negligible singular values in the normalized spectrum.

We first normalize singular values to the unit interval:

$$\tilde{\sigma}_k = \frac{\sigma_k - \sigma_{\min}}{\sigma_{\max} - \sigma_{\min}}. \tag{6}$$

Tail redundancy is then defined as the fraction of singular values whose normalized magnitude falls below a threshold τ:

$$T_i = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}[\tilde{\sigma}_k < \tau], \tag{7}$$

where $\mathbf{1}[\cdot]$ denotes the indicator function. A larger T_i indicates that a greater proportion of spectral components lie in the tail, suggesting higher structural redundancy, whereas a smaller T_i corresponds to a more compact and information-efficient spectrum. Importantly, T_i reflects structural compressibility rather than information capacity. While we

use a threshold-based instantiation here, details on robustness and design choices are discussed in Appendix A.

To modulate the baseline allocation, we map tail redundancy to a refinement score:

$$s_i = 1 - 0.5 \cdot T_i, \tag{8}$$

where higher redundancy yields lower scores and thus stronger compression signals. Scores are clipped to a bounded interval to prevent extreme reallocations. In practice, for matrices of the same type, the corresponding scores are first averaged across layers before being normalized by their mean $\bar{s} = \frac{1}{L}\sum_{i=1}^{L} s_i$. The final rank assignment is obtained as

$$r_i^{(Stage2)} = r_i^{(Stage1)} \left(1 + \beta \cdot \log\left(\frac{s_i}{\bar{s}}\right)\right), \tag{9}$$

where β controls the magnitude of redundancy-aware refinement.

By sequentially applying capacity-anchored allocation and redundancy-aware refinement, HiSVD decouples information capacity from structural redundancy, enabling systematic compression of structurally inflated layers. After the two-stage refinement, a global rank rescaling is applied to exactly satisfy the target compression ratio (see Algorithm 2 for details).

The computational overhead of HiSVD's rank allocation is negligible compared to the SVD factorization shared by all SVD-based methods, adding only $O(L)$ operations for L layers.

3.3 Post-Compression Fine-Tuning

Following prior work (Wang et al., 2024; Solgi et al., 2025), we apply LoRA-based fine-tuning (Hu et al., 2022) to recover performance after compression. This fine-tuning procedure is orthogonal to the rank allocation method and is applied uniformly across HiSVD and all baselines to ensure fair comparison.

4 Experiments and Analysis

Baselines. We benchmark HiSVD against three categories of state-of-the-art methods: (1) SVD-based Compression, including FWSVD (Hsu et al., 2022), ASVD (Yuan et al., 2023), SVD-LLM (Wang et al., 2024), Dobi-SVD (Wang et al., 2025), and AdaSVD (Li et al., 2025); (2) Structured Pruning, namely LLM-Pruner (Ma et al., 2023), SliceGPT (Ashkboos et al., 2024), and BlockPruner (Zhong et al., 2025); and (3) Quantization Synergy, comparing HiSVD combined with 4-bit quantization against SVD-LLM and the GPTQ baseline (Frantar et al., 2022).

Models and Datasets. We evaluate on five representative LLMs: LLaMA-7B (Touvron et al., 2023a), LLaMA-2-7B (Touvron et al., 2023b), LLaMA-3-8B (Grattafiori et al., 2024), OPT-6.7B (Zhang et al., 2022), and Mistral-7B (Jiang et al., 2023). Evaluation is conducted in a zero-shot setting using LM-Evaluation-Harness on 8 datasets: two for language modeling (WikiText-2 (Merity et al., 2016), C4 (Raffel et al., 2020)) and six for commonsense reasoning (OpenbookQA (Mihaylov et al., 2018), ARC-easy (Clark et al., 2018), Hellaswag (Zellers et al., 2019), PIQA (Bisk et al., 2020), MathQA (Amini et al., 2019), and WinoGrande (Sakaguchi et al., 2019)).

Experimental Configuration. Following the protocol of SVD-LLM (Wang et al., 2024), we use 256 randomly sampled sequences from WikiText-2 for calibration. All experiments are conducted on NVIDIA H20 GPUs.

4.1 Comparative Analysis

We evaluate HiSVD across six dimensions to demonstrate its efficacy and versatility.

Main Results. We benchmark LLaMA-7B on eight datasets across 20%–60% compression ratios (Table 1). HiSVD consistently outperforms baselines in both standard and fine-tuned (*) settings. Notably, our method maintains high resilience even without weight updates, validating the effectiveness of our rank allocation strategy.

Model Generality. We extend evaluation to diverse architectures (LLaMA-2/3, OPT, Mistral) and scales (up to 30B) at 20% compression. As shown in Table 2, HiSVD consistently outperforms baselines across all settings, verifying the strong generalization capability of our hierarchical approach.

SVD Competitors. Table 3 compares HiSVD against recent adaptive SVD frameworks. Our method achieves superior accuracy, demonstrating that our hierarchical allocation strategy identifies compression opportunities more effectively than the heuristics employed by competitors.

Pruning Comparison. We contextualize the performance of HiSVD against state-of-the-art structured pruning methods, benchmarking LLaMA-7B at 20% and 40% compression ratios. As shown in Table 4, HiSVD consistently achieves lower perplexity than these sparsity-based approaches. This

Ratio	Method	WikiText-2↓	C4↓	OpenB.↑	ARC_E↑	HellaS.↑	PIQA↑	MathQA↑	WinoG.↑	Average↑
0%	Original	5.68	7.34	0.28	0.67	0.56	0.78	0.27	0.67	0.54
20%	FWSVD	1727	1511	0.15	0.31	0.26	0.56	0.21	0.50	0.33
	ASVD	11.14	15.93	0.25	0.53	0.41	0.68	0.24	0.64	0.46
	SVD-LLM	7.94	15.84	0.22	0.58	0.43	0.69	0.24	0.63	0.47
	SVD-LLM*	7.73	12.23	0.33	0.67	0.55	0.79	0.26	0.69	0.55
	HiSVD	**7.33**	**13.79**	**0.27**	**0.67**	**0.45**	**0.72**	0.23	**0.67**	**0.50**
	HiSVD*	**7.14**	**11.05**	**0.33**	**0.70**	0.54	0.77	0.25	**0.70**	**0.55**
30%	FWSVD	20127	7240	0.17	0.26	0.26	0.51	0.19	0.49	0.31
	ASVD	51	41	0.18	0.43	0.37	0.65	0.21	0.53	0.40
	SVD-LLM	9.56	25.11	0.2	0.48	0.37	0.65	0.22	0.59	0.42
	SVD-LLM*	8.13	12.95	0.26	0.68	0.47	0.71	0.24	0.64	0.50
	HiSVD	**8.53**	**20.37**	**0.25**	**0.59**	**0.40**	**0.68**	**0.23**	**0.65**	**0.47**
	HiSVD*	**7.77**	**12.48**	**0.31**	**0.66**	**0.51**	**0.75**	**0.24**	**0.66**	**0.52**
40%	FWSVD	18156	12847	0.16	0.26	0.26	0.53	0.21	0.51	0.32
	ASVD	1407	1109	0.13	0.28	0.26	0.55	0.19	0.48	0.32
	SVD-LLM	13.73	75.42	0.25	0.33	0.40	0.63	0.12	0.55	0.38
	SVD-LLM*	9.27	15.63	0.29	0.59	0.52	0.69	0.20	0.68	0.49
	HiSVD	**11.48**	**38.83**	0.20	**0.49**	0.34	**0.64**	**0.22**	**0.59**	**0.41**
	HiSVD*	**8.87**	**14.61**	**0.29**	**0.63**	0.48	**0.72**	**0.24**	0.62	**0.50**
50%	FWSVD	24391	23104	0.12	0.26	0.26	0.53	0.20	0.50	0.31
	ASVD	15358	27929	0.12	0.26	0.26	0.52	0.19	0.51	0.31
	SVD-LLM	23.97	118.57	0.16	0.33	0.29	0.56	0.21	0.54	0.35
	SVD-LLM*	15.30	19.26	0.22	0.54	0.40	0.67	0.23	0.59	0.44
	HiSVD	**19.93**	**90.46**	**0.19**	**0.39**	**0.30**	**0.58**	**0.21**	**0.56**	**0.37**
	HiSVD*	**10.70**	**17.36**	**0.25**	**0.59**	**0.43**	**0.70**	**0.24**	**0.60**	**0.47**
60%[1]	FWSVD	32194	29292	0.15	0.26	0.26	0.53	0.18	0.50	0.31
	ASVD	57057	43036	0.12	0.26	0.26	0.51	0.18	0.49	0.30
	SVD-LLM	66.62	471.83	0.10	0.05	0.10	0.21	0.04	0.17	0.11
	SVD-LLM*	15.00	26.26	0.18	0.42	0.31	0.35	0.12	0.44	0.30
	HiSVD	**43.02**	**238.69**	**0.14**	**0.30**	**0.28**	**0.55**	**0.22**	**0.53**	**0.34**
	HiSVD*	**14.25**	**22.84**	**0.22**	**0.51**	**0.38**	**0.67**	**0.23**	**0.57**	**0.43**

Table 1: Performance of Llama-7B compressed by HiSVD and baselines under 20% to 60% compression ratio in terms of perplexity (on two language modeling datasets) and accuracy (on seven common sense reasoning datasets).

Method	Llama 30B	Llama-2 7B	Llama-3 8B	OPT-6.7B	Mistral-7B
Original	4.10	5.47	6.14	10.86	5.25
FWSVD	20.54	2360	4782	14559	6357
ASVD	22.71	10.10	17.55	82.00	13.72
SVD-LLM	5.63	8.50	14.41	16.04	10.21
SVD-LLM*	5.14	7.73	11.41	14.47	7.47
HiSVD	**5.19**	**7.73**	**11.77**	**11.58**	**7.02**
HiSVD*	**4.35**	**7.02**	**10.14**	**11.16**	**6.63**

Table 2: Perplexity of LLAMA-13B, LLAMA-30B, LLAMA-2-7B, LLAMA-3-8B, OPT-6.7B and Mistral-7B under 20% compression ratio.

performance differential substantiates the intrinsic superiority of low-rank approximation over structured sparsity. By explicitly isolating the principal energetic components, HiSVD ensures maximal representational integrity while upholding mathematical interpretability—attributes frequently compromised by the heuristic-driven masking inherent to pruning baselines.

Quantization Synergy. We demonstrate that HiSVD is orthogonal to quantization by integrating it with GPTQ under extreme memory budgets (Table 5). This result validates the synergy of the "low-rank + low-bit" approach, suggesting that such hybrid methodologies are pivotal for advancing the frontiers of post-training compression.

Hardware Efficiency. We measure LLaMA-7B generation throughput on an NVIDIA H20 GPU (Figure 4). HiSVD yields consistent speedups that scale with the compression ratio (20%–60%), effectively translating reduced complexity into practical acceleration.

4.2 Ablation Study

We evaluate four variants on LLaMA-7B: HiSVD-1 (Stage 1 only), HiSVD-2 (Stage 2 only), HiSVD-3 (Stage 1+2 Synergy), and HiSVD-4 (HiSVD-

[1]Extremely large perplexity values indicate collapse under aggressive compression ratio.

Method	Perplexity↓	Accuracy↑
Original	5.68	0.53
Ada-SVD	14.76	0.34
Dobi-SVD	13.54	0.38
HiSVD	**11.48**	**0.41**
HiSVD*	**8.87**	**0.50**

Table 3: Compared to Dobi-SVD and AdaSVD on Llama-7B at 40% Compression Ratio.

Ratio	Model	WikiText2
20%	LLM-Pruner	9.88
	SliceGPT	8.78
	BlockPruner	9.4
	HiSVD	**7.14**
40%	LLM-Pruner	18.94
	SliceGPT	16.39
	BlockPruner	19.78
	HiSVD	**8.87**

Table 4: Perplexity of LLAMA-7B compressed by structured pruning methods and HiSVD.

Method	Memory	Perplexity↓
GPTQ-3bit	2.8GB	16.28
SVD-LLM	2.8GB	35.98
SVD-LLM+GPTQ-4bit	2.8GB	10.36
HiSVD	2.8GB	**32.35**
HiSVD+GPTQ-4bit	2.8GB	**9.69**

Table 5: Perplexity of LLAMA-7B compressed by GPTQ and HiSVD.

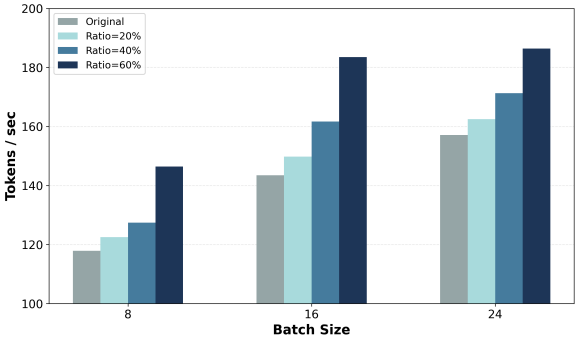

Figure 4: Throughput (Tokens/sec) of LLAMA-7B and its compressed version by HiSVD on single H20 GPU.

	20% Compression Ratio			40% Compression Ratio			60% Compression Ratio		
Method	WikiText2↓	C4↓	Accuracy↑	WikiText2↓	C4↓	Accuracy↑	WikiText2↓	C4↓	Accuracy↑
SVD-LLM	7.94	15.84	0.47	13.73	75.42	0.38	66.62	471.83	0.11
SVD-LLM*	7.73	12.23	0.55	9.27	15.63	0.49	15.00	26.26	0.30
HiSVD-1	**7.47**	**14.49**	**0.49**	**12.02**	**43.41**	**0.41**	**49.21**	**294.34**	**0.34**
HiSVD-2	**7.40**	**14.09**	**0.50**	**11.66**	**41.05**	**0.41**	**45.34**	**265.67**	**0.34**
HiSVD-3	**7.33**	**13.79**	**0.50**	**11.48**	**38.83**	**0.41**	**43.02**	**238.69**	**0.34**
HiSVD-4	**7.14**	**11.05**	**0.55**	**8.87**	**14.61**	**0.50**	**14.25**	**22.84**	**0.43**

Table 6: Ablation study on LLaMA-7B across varying compression ratios. **HiSVD-1**: Stage 1 only; **HiSVD-2**: Stage 2 only; **HiSVD-3**: Stage 1+2 Synergy; **HiSVD-4**: HiSVD-3 with fine-tuning.

3 with fine-tuning). Table 6 shows that HiSVD-3 consistently outperforms single-stage baselines. This confirms that while Stage 1 captures global capacity trends, Stage 2 exploits the "imperfect coupling" by targeting tail redundancy that Stage 1 overlooks. Moreover, the robust recovery of HiSVD-4 validates that our allocation preserves the "spectral skeleton" even before weight updates. Unlike opaque heuristics, these results demonstrate HiSVD's interpretability by disentangling information capacity from structural redundancy.

4.3 Robustness and Design Choices

To further examine the stability of HiSVD, we evaluate the sensitivity of the redundancy threshold τ used in Stage II. Specifically, we vary $\tau \in \{0.05, 0.10, 0.15, 0.20, 0.25\}$ while keeping all other settings fixed. As shown in Table 7, the resulting fluctuation is extremely small, with the variance on WikiText-2 below 0.0004. Although $\tau = 0.15$ achieves marginally better performance, we use $\tau = 0.10$ as the default setting throughout the paper because it provides a highly robust and effective operating point.

This stability stems from the redundancy-aware refinement mechanism itself. Even if the initial threshold slightly overestimates or underestimates the tail region, Stage II dynamically rebalances the final rank allocation according to the observed spectral redundancy. As a result, HiSVD is insensitive to the precise value of τ, provided it lies within a reasonable range.

We also examine the robustness of the scaling hyperparameters α and β. Our default setting is $\alpha = 1$ and $\beta = 1$, which is used throughout the main experiments without additional tuning. As shown in

τ	0.05	0.10(default)	0.15	0.20	0.25
WikiText2	7.76	7.73	7.71	7.73	7.74
C4	16.89	16.58	16.39	16.51	16.65

Table 7: Sensitivity of HiSVD to the redundancy threshold τ on LLAMA-2-7B under 20% compression ratio.

	$\alpha = 0.5, \beta = 1$	$\alpha = 1, \beta = 1$(default)	$\alpha = 1.5, \beta = 1$	$\alpha = 1, \beta = 0.5$	$\alpha = 1, \beta = 1.5$
WikiText2	7.74	7.73	7.75	7.81	7.72
C4	16.58	16.58	16.59	17.00	16.24

Table 8: Sensitivity of HiSVD to the scaling hyperparameters α and β on LLAMA-2-7B under 20% compression ratio.

Table 8, varying either parameter within $[0.5, 1.5]$ leads to only minor changes in perplexity. In particular, the fluctuation on WikiText-2 remains below 0.04 PPL, confirming that $(\alpha, \beta) = (1, 1)$ is a stable operating point rather than a carefully tuned configuration.

From a conceptual perspective, α and β act as intensity control knobs with clear physical interpretations. The parameter α controls the sensitivity to information capacity in Stage I, while β regulates the aggressiveness of the redundancy-aware refinement in Stage II. We observe that changing α to either 0.5 or 1.5 results in only slight degradation, indicating that the capacity-anchored allocation is inherently robust. In contrast, reducing β to 0.5 noticeably worsens performance, demonstrating that weakening the redundancy penalty prevents HiSVD from fully exploiting structural redundancy. This empirically validates the necessity of Stage II. Interestingly, increasing β to 1.5 yields slightly better results in some cases, suggesting that the model contains substantial compressible redundancy. Nevertheless, we adopt $\beta = 1$ as a balanced and safe default that remains effective across diverse architectures and compression ratios.

The design choices in Equation 8 are motivated by the need to balance the two complementary factors captured by HiSVD: information capacity and spectral redundancy. In particular, the constant 0.5 acts as a damping factor that ensures stable and conservative rank refinement. The raw Tail Redundancy score measures the proportion of negligible singular values in each layer. If it were applied directly with a scaling factor of 1, the refinement could become overly aggressive for highly redundant layers, potentially leading to representational collapse. Using 0.5 provides a safety margin, ensuring that even layers with large redundancy scores retain at least 50% of their baseline rank allocation. This conservative design balances the benefit of exploiting spectral redundancy against the risk of excessive information loss. The coefficient 0.5 is introduced to moderate the magnitude of the redundancy correction relative to the effective-rank baseline, preventing Stage II from overreacting to noisy tail singular values.

5 Conclusion

In this work, we identify a fundamental limitation in SVD-based compression: the widespread reliance on metrics that fail to distinguish between representational importance and structural compressibility. While existing methods typically treat layer importance as a sufficient proxy for rank allocation, our spectral analysis reveals that this coupling is imperfect—layers with comparable information capacity often exhibit markedly different spectral decay patterns, corresponding to varying degrees of redundancy in the spectral tail. To bridge this gap, we propose HiSVD, a hierarchical framework that explicitly decouples these two factors. By integrating Capacity-Anchored Baseline Allocation to establish representational stability with Redundancy-Aware Refinement to target structural excess, HiSVD effectively exploits fine-grained compression opportunities that performance-oriented proxies overlook. Extensive experiments demonstrate that this principled decoupling enables more efficient compression and consistently outperforms state-of-the-art SVD-based methods, highlighting the necessity of modeling spectral distribution shape beyond capacity-centric abstractions.

Limitations

While HiSVD demonstrates significant efficacy in optimizing low-rank compression by theoretically decoupling information capacity from spectral distribution shape, several limitations warrant further investigation.

Computational Complexity of Spectral Decomposition. Although the rank allocation phase of HiSVD is computationally efficient ($O(L)$), the prerequisite step involves performing full Singular Value Decomposition (SVD) on every single weight matrix. This factorization process is computationally intensive, particularly for high-dimensional weight matrices found in massive-scale models. Consequently, the preprocessing time grows significantly with model size, which presents a bottleneck for extremely large models or scenarios requiring frequent re-compression. While this is a one-time offline cost, it limits the feasibility of dynamic, on-the-fly execution in resource-constrained edge environments. Future work could explore randomized SVD algorithms or iterative approximation methods to mitigate this computational overhead.

Inference Latency Characteristics. HiSVD significantly improves decoding throughput (tokens/sec) by reducing memory bandwidth usage. However, the Time-To-First-Token (TTFT) latency during the prefill phase may not see proportional gains—or may slightly regress—due to the kernel scheduling overhead introduced by replacing a single dense matrix multiplication with two sequential low-rank operations. Since HiSVD preserves the original attention structure and KV-cache layout, its impact on TTFT is expected to remain limited. Therefore, the acceleration benefits of HiSVD are most pronounced in memory-bound, steady-state decoding scenarios, while a dedicated end-to-end latency study remains an important direction for future work.

Heuristic Design and Hyperparameter Sensitivity. HiSVD relies on several design choices, including the min-max normalization for tail redundancy and the modulation coefficients (α, β). Although our experiments show that the same default configuration generalizes well across the LLaMA, OPT, and Mistral families, these heuristics may require revalidation for architectures with fundamentally different spectral signatures, such as Mixture-of-Experts (MoE) models. Nevertheless, because HiSVD computes information capacity and tail redundancy independently for each layer, the framework is naturally compatible with heterogeneous architectures and could potentially adapt to different experts without assuming structural uniformity. In addition, HiSVD currently relies solely on weight spectral information. Incorporating task-aware signals, such as activation statistics or Fisher information, may further improve the sensitivity of the allocation strategy to downstream tasks. Developing a more adaptive, parameter-free mechanism that automatically derives these coefficients and robustifies the tail metric therefore remains an important direction for future research.

Theoretical Bounds of Tail Redundancy. Our methodology leverages tail redundancy to identify structurally inflated layers, operating on the premise that singular values with negligible magnitude represent removable structural excess. While statistically valid for general language modeling performance, there is a theoretical possibility that certain heavy-tail singular values—though numerically small—may encode "long-tail" factual knowledge or rare linguistic patterns. Aggressive pruning based on spectral distribution shape might disproportionately impact the recall of such obscure facts, a phenomenon that standard perplexity metrics on broad corpora might not fully capture. Further investigation is needed to quantify the relationship between spectral tail pruning and fine-grained knowledge retention.

Dependence on Post-Compression Fine-Tuning. Consistent with state-of-the-art low-rank compression frameworks, HiSVD utilizes LoRA-based fine-tuning to compensate for the precision loss incurred by rank truncation. While our method achieves superior zero-shot performance compared to baselines (Table 1), achieving near-lossless recovery (HiSVD*) still necessitates a re-training phase with access to calibration data and gradient computation. This requirement constrains the applicability of the method in strictly "data-free" or "gradient-free" deployment scenarios where only pre-trained weights are accessible.

References

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*

(Long and Short Papers), pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Saleh Ashkboos, Maximilian L Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. 2024. Slicegpt: Compress large language models by deleting rows and columns. *arXiv preprint arXiv:2401.15024*.

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. 2020. Piqa: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457.

Nadav Cohen, Or Sharir, and Amnon Shashua. 2016. On the expressive power of deep learning: A tensor analysis. In *Conference on learning theory*, pages 698–728. PMLR.

Tim Dettmers, Ruslan Svirschevski, and 1 others. 2023. Spqr: A sparse-quantized representation for near-lossless llm weight compression. *Transactions on Machine Learning Research*. To appear.

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*.

Shangqian Gao, Ting Hua, Yen-Chang Hsu, Yilin Shen, and Hongxia Jin. 2024. Adaptive rank selections for low-rank approximation of language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 227–241, Mexico City, Mexico. Association for Computational Linguistics.

Gene H Golub, Alan Hoffman, and Gilbert W Stewart. 1987. A generalization of the eckart-young-mirsky matrix approximation theorem. *Linear Algebra and its applications*, 88:317–327.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 8003–8017.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Ajay Kumar Jaiswal, Yifan Wang, Lu Yin, Shiwei Liu, Runjin Chen, Jiawei Zhao, Ananth Grama, Yuandong Tian, and Zhangyang Wang. 2025. From low rank gradient subspace stabilization to low-rank weights: Observations, theories, and applications. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 26740–26756. PMLR.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Zhiteng Li, Mingyuan Xia, Jingyuan Zhang, Zheng Hui, Haotong Qin, Linghe Kong, Yulun Zhang, and Xiaokang Yang. 2025. Adasvd: Adaptive singular value decomposition for large language models. *arXiv preprint arXiv:2502.01403*.

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.

Charles H Martin and Michael W Mahoney. 2021. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *CoRR*, abs/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

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.

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. *J. Mach. Learn. Res.*, 21(1).

Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V. Le. 2019. Regularized evolution for image classifier architecture search. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Olivier Roy and Martin Vetterli. 2007. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pages 606–610. IEEE.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

Wenqi Shao, Mengzhao Chen, Zhaoyang Zhang, Peng Xu, Lirui Zhao, Zhiqiu Li, Yu Qiao, Ping Luo, Kaipeng Wang, and Zhang Zhang. 2024. OmniQuant: Omnidirectionally calibrated quantization for large language models. In *The Twelfth International Conference on Learning Representations*.

Ryan Solgi, Parsa Madinei, Jiayi Tian, Rupak Swaminathan, Jing Liu, Nathan Susanj, and Zheng Zhang. 2025. Activation-informed pareto-guided low-rank compression for efficient llm/vlm. *arXiv preprint arXiv:2510.05544*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. 2024. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations*.

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.

Roman Vershynin. 2018. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Qinsi Wang, Jinghan Ke, Masayoshi Tomizuka, Yiran Chen, Kurt Keutzer, and Chenfeng Xu. 2025. Dobi-svd: Differentiable svd for llm compression and some new perspectives. *arXiv preprint arXiv:2502.02723*.

Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. 2024. Svd-llm: Truncation-aware singular value decomposition for large language model compression. *arXiv preprint arXiv:2403.07378*.

Zhihang Yuan, Yuzhang Shang, Yue Song, Dawei Yang, Qiang Wu, Yan Yan, and Guangyu Sun. 2023. Asvd: Activation-aware singular value decomposition for compressing large language models. *arXiv preprint arXiv:2312.05821*.

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, Florence, Italy. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, and 1 others. 2024. A survey of large language models. *Transactions of the Association for Computational Linguistics*, 12:55–100.

Longguang Zhong, Fanqi Wan, Ruijun Chen, Xiaojun Quan, and Liangzhi Li. 2025. Blockpruner: Fine-grained pruning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5065–5080.

A Implementation Details and Design Choices

A.1 Hyperparameters and Normalization

To demonstrate the robustness and ease of deployment of HiSVD, we employ a unified hyperparameter configuration across all experiments: $\alpha = 1$, $\beta = 1$, and $\tau = 0.1$. We explicitly did not perform per-model or per-ratio hyperparameter tuning. The fact that this single configuration consistently outperforms baselines across diverse architectures (LLaMA, OPT, Mistral) and scales (7B to 30B) validates that our hierarchical decoupling strategy captures intrinsic spectral properties rather than overfitting to specific hyperparameters.

In Stage I, the effective rank is computed as a normalized quantity by dividing the raw participation-ratio-style effective rank by the corresponding matrix dimensionality. In Stage II, the tail redundancy score is likewise computed from normalized singular values. The zero-mean adjustment is applied to these normalized scores to ensure relative comparability across layers.

A.2 Budget Control via Global Rank Rescaling

Equation 1 is used to compute the baseline rank allocation that corresponds to the target compression ratio. The subsequent Stage I and Stage II adjustments operate on these baseline ranks to redistribute capacity across layers according to normalized capacity and redundancy signals.

As these two stages are designed to perform relative reallocation, the intermediate rank assignments may temporarily deviate from the original target compression ratio. To ensure that the final model strictly matches the desired global compression level, we apply a global rank rescaling step after Stage II.

Specifically, all layer-wise ranks are uniformly scaled by a constant factor so that the total number of retained parameters exactly matches the target compression ratio specified by Eq. 1. The final scaled ranks are rounded to the nearest integers to obtain valid matrix dimensions. This process preserves the relative allocation pattern induced by HiSVD while strictly enforcing the global compression constraint.

A.3 Justification for Effective Rank Proxy

In Stage I, we adopt the effective rank definition $R_{eff} = (\sum \sigma_k)^2 / \sum \sigma_k^2$. While entropy-based effective rank and stable rank are valid alternatives for measuring dimensionality, we select this specific formulation because it acts as a normalized proxy for the "participation ratio" of singular values. It provides a computationally efficient and scale-invariant measure of spectral concentration without requiring logarithmic operations. Our focus is on the relative distribution of capacity across layers rather than the absolute theoretical value of the rank; empirically, this metric correlates well with the information capacity required for the baseline allocation.

A.4 Robustness of Tail Redundancy

In Stage II, we characterize spectral compressibility using a tail redundancy score derived from min-max normalized singular values. We acknowledge that min-max normalization can be sensitive to outliers in the spectral distribution. However, in the context of pre-trained LLMs, we observe that weight spectra are generally well-behaved, and the specific threshold τ serves as a consistent control variate across layers. While alternative measures (e.g., quantile-based tail mass or power-law fitting) could potentially offer greater theoretical robustness, our specific instantiation provides a direct and interpretable signal for identifying structural inflation, which is the primary goal of the refinement stage.

A.5 Motivation for Redundancy-Aware Refinement

The logarithmic refinement in Stage II is motivated by the intuition from the Eckart-Young-Mirsky theorem, which states that the approximation error is determined by the sum of squared tail singular values. Layers with "heavy tails" (high redundancy score s_i) imply that a larger portion of the spectral energy is concentrated in smaller components, allowing for more aggressive truncation without significant information loss. The logarithmic modulation function is chosen to provide a bounded, non-linear adjustment that penalizes structural excess while preventing drastic rank fluctuations that could destabilize the network. A formal theoretical justification, demonstrating how this refinement guarantees a first-order reduction in reconstruction error under budget constraints, is provided in Appendix B.

A.6 AI Assistant Declaration

We have utilized a Large Language Model (LLM) to assist in refining the language and improving the clarity of the manuscript. The LLM was primarily used for enhancing the overall phrasing and readability of the text, ensuring a more polished and professional presentation.

B Theoretical Interpretation of HiSVD

We clarify that "decoupling" in HiSVD refers to an operational separation of functional roles, rather than statistical orthogonality between proxies. Our two-stage design assigns information capacity and spectral tail redundancy to distinct stages of rank allocation, without assuming these signals to be uncorrelated. Empirically, they exhibit strong negative correlation (Figure 1), which motivates a hierarchical treatment: perfect alignment would render refinement redundant, while independence would obviate staging. HiSVD operates in this moderately coupled regime.

Setup and Two-Stage Allocation. Let $W_i \in \mathbb{R}^{m_i \times n_i}$ be a layer weight matrix with singular values $\sigma_{i,1} \geq \cdots \geq \sigma_{i,d_i} \geq 0$, where $d_i = \min(m_i, n_i)$. The truncated-SVD reconstruction error at rank r is

$$E_i(r) = \|W_i - W_{i,r}\|_F^2 = \sum_{j>r} \sigma_{i,j}^2. \tag{10}$$

We use the normalized information capacity $\rho_i \in (0, 1]$ defined in Eq. 4 as the primary anchor. Recalling Eq. 5, Stage I (*capacity-anchored baseline allocation*) determines the rank $r_i^{(\text{Stage1})}$ based on the uniform baseline r_{base} and the capacity deviation:

$$r_i^{(\text{Stage1})} = r_{\text{base}} \left(1 + \alpha(\rho_i - \bar{\rho})\right), \tag{11}$$

where α controls the anchoring strength and $\bar{\rho}$ is the mean capacity. Since $\sum(\rho_i - \bar{\rho}) = 0$, this stage naturally preserves the total parameter budget relative to the uniform baseline.

Let s_i denote the refinement score derived from tail redundancy (Eq. 8) and $\bar{s}$ its mean. Stage II (*redundancy-aware refinement*) applies a bounded multiplicative logarithmic adjustment:

$$r_i^{(\text{Stage2})} = r_i^{(\text{Stage1})} \left(1 + \beta \log \frac{s_i}{\bar{s}}\right). \tag{12}$$

A final rescaling factor γ restores strict budget consistency:

$$r_i = \lfloor \gamma \, r_i^{(\text{Stage2})} \rceil, \quad \text{s.t.} \quad \sum_i r_i(m_i + n_i) \approx B. \tag{13}$$

Stability and Conditional Improvement. Stage II is designed as a relative perturbation around the capacity-anchored baseline. Let the perturbation magnitude be defined as:

$$\delta = \max_i \left| \beta \log \frac{s_i}{\bar{s}} \right|. \tag{14}$$

If $\delta < 1$, then the refinement is uniformly bounded:

$$(1-\delta) r_i^{(\text{Stage1})} \leq r_i^{(\text{Stage2})} \leq (1+\delta) r_i^{(\text{Stage1})}. \tag{15}$$

Thus, Stage II modifies the rank distribution without overturning the foundational capacity structure established in Stage I. Given the definition $s_i = 1 - 0.5 \cdot T_i$ (Eq. 8), the refinement score is strictly bounded within $s_i \in (0.5, 1)$. With the default setting $\beta = 1$, the maximum perturbation magnitude is bounded by $\delta \leq \ln 2 \approx 0.693$. Consequently, the stability condition $\delta < 1$ is naturally satisfied by design, guaranteeing non-negative ranks without requiring additional explicit clipping. Consider a generic local rank perturbation Δr. The change in reconstruction error is:

$$E_i(r) - E_i(r + \Delta r) = \sum_{j=r+1}^{r+\Delta r} \sigma_{i,j}^2 \approx \Delta r \cdot \sigma_{i,r+1}^2. \tag{16}$$

By applying this approximation to the specific rank shift $\Delta_i = r_i^{(\text{Stage2})} - r_i^{(\text{Stage1})}$, the first-order linearization yields:

$$E_i(r_i^{(\text{Stage2})}) \approx E_i(r_i^{(\text{Stage1})}) - \Delta_i \cdot \sigma_{i, r_i^{(\text{Stage1})}+1}^2. \tag{17}$$

Proposition (Conditional First-Order Decrease). To analyze the aggregate error trend, we consider the budget constraint in a parameter-normalized context. This simplifies the strict parameter conservation to a rank-sum constraint $\sum_i \Delta_i \approx 0$. Under this approximation, the aggregate change in error is:

$$\sum_i \Delta E_i \approx - \sum_i \Delta_i \cdot \sigma_{i, r_i^{(\text{Stage1})}+1}^2. \tag{18}$$

Since Stage II increases rank ($\Delta_i > 0$) for layers with low redundancy (high s_i)—which typically possess larger marginal singular values

$\sigma^2_{i,r+1}$—and decreases rank for high-redundancy layers, the inner product is negative:

$$\sum_i \left(E_i(r_i^{(\text{Stage2})}) - E_i(r_i^{(\text{Stage1})})\right) < 0. \quad (19)$$

This provides a local guarantee: anchoring on capacity and refining towards layers with larger marginal singular energy reduces the aggregate reconstruction loss. The validity of Eq. (19) relies on the positive correlation between the refinement score s_i and the marginal spectral energy $\sigma^2_{i,r+1}$. Intuitively, layers with low tail redundancy (high s_i) typically exhibit a slower decay in their singular value spectrum, implying a larger $\sigma^2_{i,r+1}$. By shifting the budget $\Delta_i > 0$ to these layers, HiSVD maximizes the gradient of error reduction ($\Delta_i \cdot \sigma^2$), ensuring the aggregate reconstruction loss decreases.

Remarks. These results are intentionally local and do not assert global optimality. Their role is to formalize (i) Stage I as a capacity-anchored reference, (ii) Stage II as a conservative logarithmic refinement, and (iii) a verifiable condition under which spectral-aware refinement reduces error.

C Extended Spectral Analysis Across Architectures

We extend the analysis of imperfect coupling (Figure 1) to a broader range of architectures, covering varying scales (LLaMA-2-7B, LLaMA-30B) and distinct attention mechanisms (Grouped Query Attention in LLaMA-3-8B and Mistral-7B).

C.1 Metric Consistency and Universality

Crucially, our analysis relies on the dimension-normalized information capacity (ρ_i). As discussed in Section 3, this normalization is pivotal for comparing heterogeneous layers. It effectively prevents the reduced-dimension Key/Value projections in GQA models from appearing as statistical outliers, ensuring they share a common basis with larger Query/Output and MLP layers.

C.2 Analysis Results

Figures 5 through 9 present the scatter plots for both MHA and GQA architectures. By employing ρ_i, we observe a striking universality:

- **Consistency Across Scales (MHA):** Across LLaMA-7B, LLaMA-2-7B, and LLaMA-30B, the capacity-redundancy trade-off remains robust to model scaling, indicating the distribution shape is intrinsic to the learning process rather than model size.
- **Alignment of GQA Structures:** For LLaMA-3-8B and Mistral-7B, ρ_i successfully aligns the structurally distinct GQA Key/Value layers with the global trend. Despite having significantly smaller n_i, these layers integrate seamlessly with Q/O and MLP layers, avoiding the clustering artifacts seen with raw rank metrics.

These results confirm that imperfect coupling is a fundamental property of pre-trained Transformer weights, rather than an artifact of specific dimensions or attention mechanisms. Furthermore, the consistent post-compression alignment ($R^2 > 0.89$) validates that HiSVD's hierarchical strategy naturally adapts to these structural variations without requiring architecture-specific heuristics.

D Visualization of Hierarchical Rank Allocation

We visualize the proposed two-stage rank allocation on LLaMA-7B, illustrating how the budget is dynamically distributed across layers and submodules (e.g., `att.q`, `mlp.up`) based on intrinsic properties.

D.1 Stage I: Capacity-Anchored Baseline Allocation

Figure 10 visualizes the rank allocation process in Stage I.

- **Metric Curve (ρ):** The purple curve floating above represents the estimated Capacity (ρ) for each module across different layers. A higher value signifies a higher capacity requirement to capture complex feature transformations.
- **Allocation Bars (Δr):** On the projection plane, the black line serves as the reference axis representing the Uniform Baseline (i.e., $\Delta r = 0$). The colored bars projecting from this axis indicate the deviation in rank allocation:
  - **Blue Bars (Rank Increase):** For modules with high capacity scores (e.g., MLP layers), the algorithm allocates additional rank budget ($\Delta r > 0$).

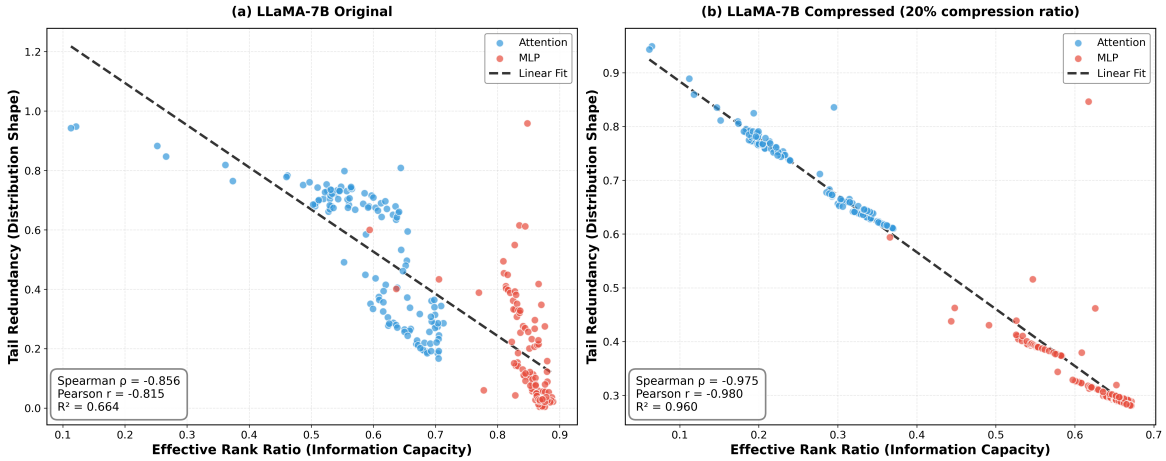

Figure 5: Scatter plot comparison for LLaMA-7B (MHA). (a) Original model showing the characteristic negative correlation. (b) HiSVD significantly tightens this coupling after 20% compression.

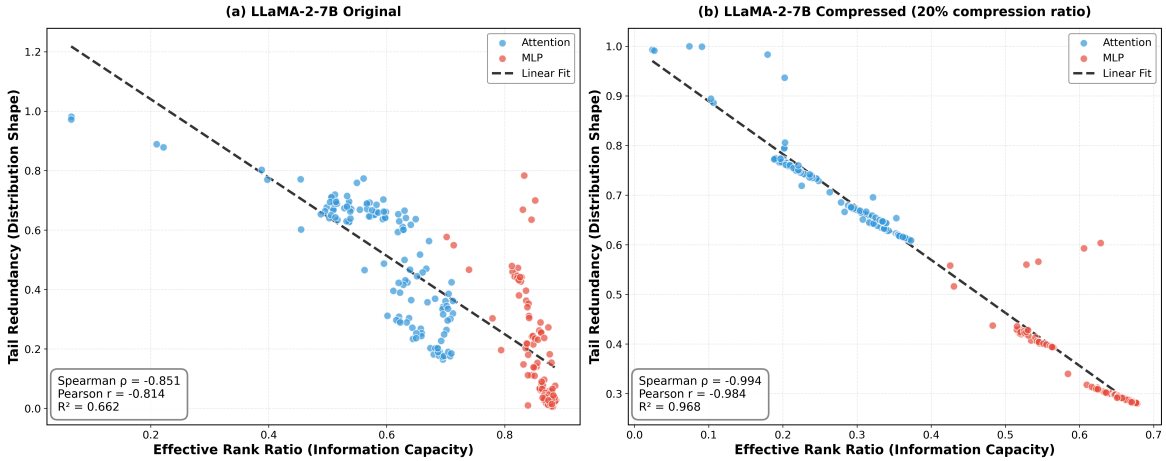

Figure 6: Scatter plot comparison for LLaMA-2-7B (MHA). The spectral behavior remains consistent across LLaMA generations.

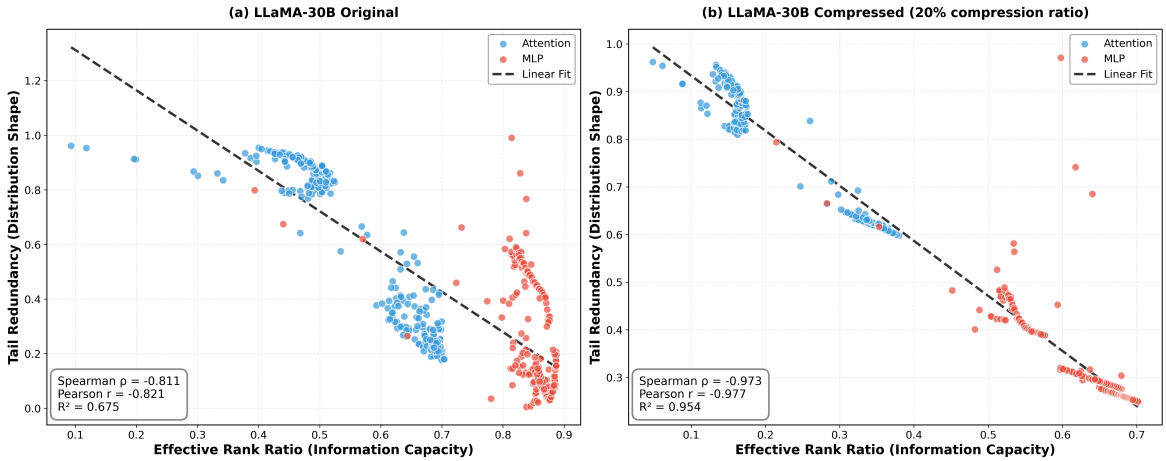

Figure 7: Scatter plot comparison for LLaMA-30B (MHA). Despite the massive increase in parameter count, the normalized spectral distribution follows the same trajectory.

– **Red Bars (Rank Decrease):** For modules with lower capacity scores (e.g., `att.q`, `att.k`), the rank is reduced ($\Delta r < 0$) relative to the uniform base-

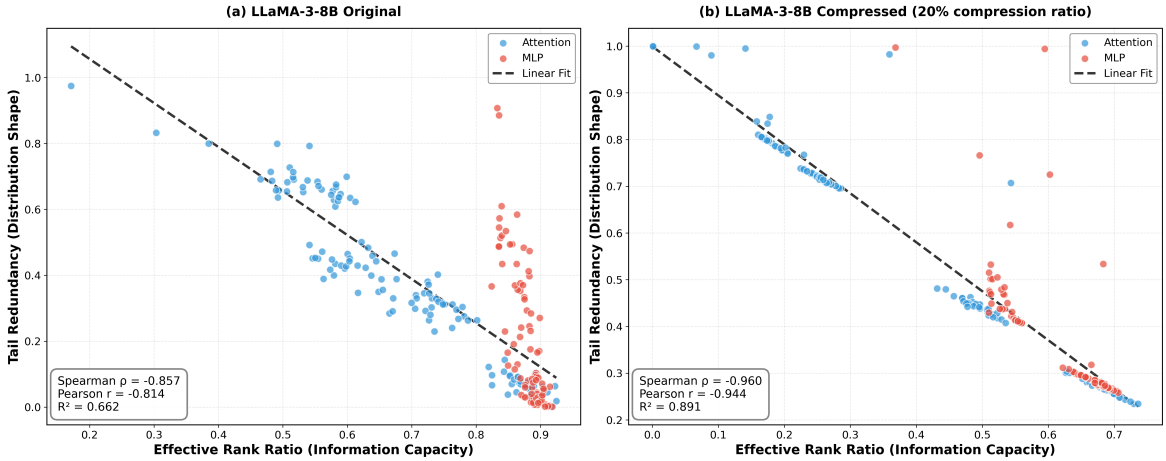

Figure 8: Scatter plot comparison for LLaMA-3-8B (GQA). By normalizing effective rank by layer dimension ($\min(m, n)$), the structurally distinct K/V layers align perfectly with the global spectral trend, demonstrating that the imperfect coupling is an intrinsic property independent of architectural variations.

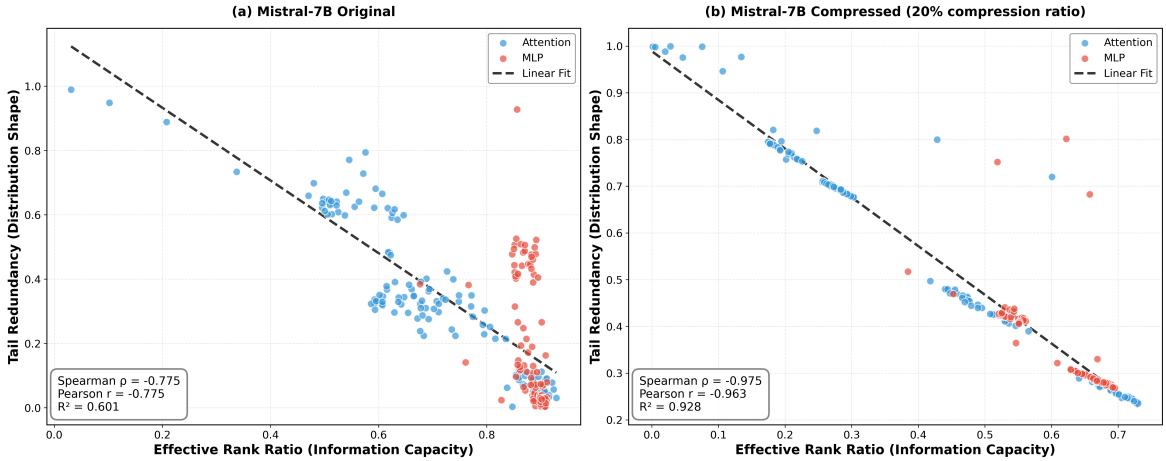

Figure 9: Scatter plot comparison for Mistral-7B (GQA). Consistent with LLaMA-3-8B, HiSVD effectively exploits the spectral heterogeneity across all layer types.

line.

D.2 Stage II: Redundancy-Aware Refinement

Figure 11 demonstrates the refinement process in Stage II.

- **Metric Curve (τ):** The purple curve represents the Tail Redundancy (τ), derived from the singular value spectrum. A higher value implies higher redundancy in the weight updates.

- **Refinement Bars (Δr):** The black reference line on the projection plane now represents the rank allocation result from Stage 1. The bars illustrate the secondary adjustment:

  – **Red Bars (Further Compression):** Where the redundancy τ is high (e.g., parts of `att.q`), the algorithm further reduces the rank to eliminate redundancy.

  – **Blue Bars (Rank Restoration):** Conversely, where redundancy is low (implying rich spectral information), the algorithm increases the rank relative to Stage 1 to preserve information, acting as a fine-grained correction.

Overall, these visualizations empirically validate the significant heterogeneity of parameter importance across different layers and modules. By synergizing the macroscopic capacity estimation in Stage 1 with the microscopic spectral refinement in Stage 2, our approach effectively captures these distinctions. The rank budget is dynamically directed towards information-rich components (e.g., deep

MLP layers) while further compressing redundant ones (e.g., Attention Q/K projections). This ensures that limited resources are utilized where they contribute most to model expressiveness, offering a clear advantage over uniform allocation baselines.

E Algorithm

Algorithm 1 outlines the HiSVD pipeline. It begins by pre-computing and caching whitened spectral components using calibration data C. The framework then delegates rank determination to Algorithm 2, which computes layer-specific budgets $\mathcal{R}$ via a two-stage capacity-redundancy analysis and global budget enforcement. Finally, the compressed model is reconstructed using the cached matrices with inverse whitening and optimized via LoRA fine-tuning.

Algorithm 1 HiSVD Framework

Input: Original LLM M, Calibration data C
Output: Compressed LLM M'
procedure HiSVD(M)
  Initialize $\mathcal{S}_{whitening}, \mathcal{S}_{\sigma}, \mathcal{S}_{svd} \leftarrow \emptyset$
  $\mathcal{W} \leftarrow$ linear layers in M
  # Phase 1: Whitening & Decomposition
  for W in $\mathcal{W}$ **do**
    $X \leftarrow M(W, C)$
    $S \leftarrow \text{Cholesky}(XX^{\top})$
    $U, \Sigma, V^{\top} \leftarrow \text{SVD}(W \cdot S)$
    $\mathcal{S}_{whitening}$.add(S)
    $\mathcal{S}_{\sigma}$.add(Σ)
    $\mathcal{S}_{svd}$.add($U, \Sigma, V^{\top}$)
  end for
  # Phase 2: Rank Allocation (See Alg. 2)
  $\mathcal{R} \leftarrow$ **Hierarchical Rank Allocation**($\mathcal{S}_{\sigma}$)
  # Phase 3: Truncation & Update
  $i \leftarrow 1$
  for W in $\mathcal{W}$ **do**
    $S \leftarrow \mathcal{S}_{whitening}[i]$
    $U, \Sigma, V^{\top} \leftarrow \mathcal{S}_{svd}[i]$
    $r \leftarrow \mathcal{R}[i]$
    $U_r, \Sigma_r, V_r^{\top} \leftarrow U_{:,:r}, \Sigma_{:r,:r}, V^{\top}_{:r,:}$
    $W_u \leftarrow U_r \Sigma_r$
    $W_v \leftarrow V_r^{\top} S^{-1}$
    Replace W with $W_u \cdot W_v$ in M
    $i \leftarrow i + 1$
  end for
  # Phase 4: Recovery
  $M' \leftarrow$ LoRA-Finetuning(M, C)
  return M'
end procedure

Algorithm 2 Hierarchical Rank Allocation

Input: Singular values $\mathcal{S}_{\sigma}$, Params $ratio_{target}, \alpha, \beta, \tau$
Output: Final rank list $\mathcal{R}$
procedure Hierarchical Rank Allocation($\mathcal{S}_{\sigma}$)
  Initialize $\mathcal{S}_{\rho}, \mathcal{S}_{s} \leftarrow \emptyset$
  # Step 1: Spectral Analysis
  for Σ in $\mathcal{S}_{\sigma}$ **do**
    $R_{\text{eff}} \leftarrow (\sum \sigma_k)^2 / \sum \sigma_k^2$
    $\rho \leftarrow R_{\text{eff}} / \min(m, n)$ {Normalized Capacity}
    $\tilde{\sigma}_k \leftarrow (\sigma_k - \sigma_{min})/(\sigma_{max} - \sigma_{min})$
    $T \leftarrow \frac{1}{\min(m,n)} \sum \mathbf{1}[\tilde{\sigma}_k < \tau]$ {Tail Redundancy}
    $s \leftarrow 1 - 0.5 \cdot T$
    $\mathcal{S}_{\rho}$.add(ρ)
    $\mathcal{S}_{s}$.add(s)
  end for
  # Step 2: Capacity-Anchored Baseline Allocation
  $\bar{\rho} \leftarrow \text{Mean}(\mathcal{S}_{\rho})$
  $\mathcal{R}_{temp} \leftarrow \emptyset$
  for ρ in $\mathcal{S}_{\rho}$ **do**
    $r_{base} \leftarrow \left\lfloor \frac{(m \cdot n \cdot ratio_{target})}{(m+n)} \right\rfloor$
    $r^{(Stage1)} \leftarrow r_{base} \cdot (1 + \alpha(\rho - \bar{\rho}))$
    $\mathcal{R}_{temp}$.add($r^{(Stage1)}$)
  end for
  # Step 3: Redundancy-Aware Refinement
  $\bar{s} \leftarrow \text{Mean}(\mathcal{S}_s)$, where S_s is implicitly grouped by matrix type and statistics are computed within each group in practice.
  $\mathcal{R}_{temp2} \leftarrow \emptyset$
  for $s, r^{(Stage1)}$ in $(\mathcal{S}_s, \mathcal{R}_{temp1})$ **do**
    $r^{(Stage2)} \leftarrow r^{(Stage1)} \cdot (1 + \beta \log(\frac{s}{\bar{s}}))$
    $\mathcal{R}_{temp2}$.add($r^{(Stage2)}$)
  end for
  # Step 4: Global Budget Rescaling
  $P_{target} \leftarrow \sum(m_i \cdot n_i) \cdot ratio_{target}$
  $P_{current} \leftarrow \sum r_i^{(Stage2)} \cdot (m_i + n_i)$
  $\gamma \leftarrow P_{target}/P_{current}$ *{Scaling Factor}*
  $\mathcal{R} \leftarrow \emptyset$
  for r in $\mathcal{R}_{temp2}$ **do**
    $\mathcal{R}$.add(Round($r \cdot \gamma$))
  end for
  return $\mathcal{R}$
end procedure

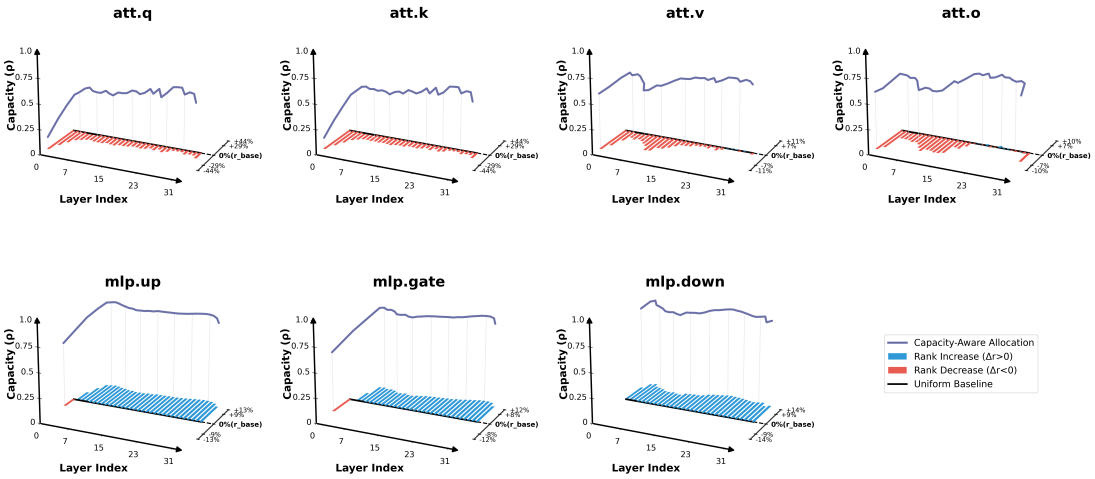

Figure 10: Visualization of Stage I (Capacity-Anchored Baseline Allocation). The purple curve indicates the Module Capacity (ρ). The black line on the projection plane acts as the Uniform Baseline reference. Blue bars denote a rank increase for high-capacity modules, while Red bars denote a rank decrease for low-capacity modules.

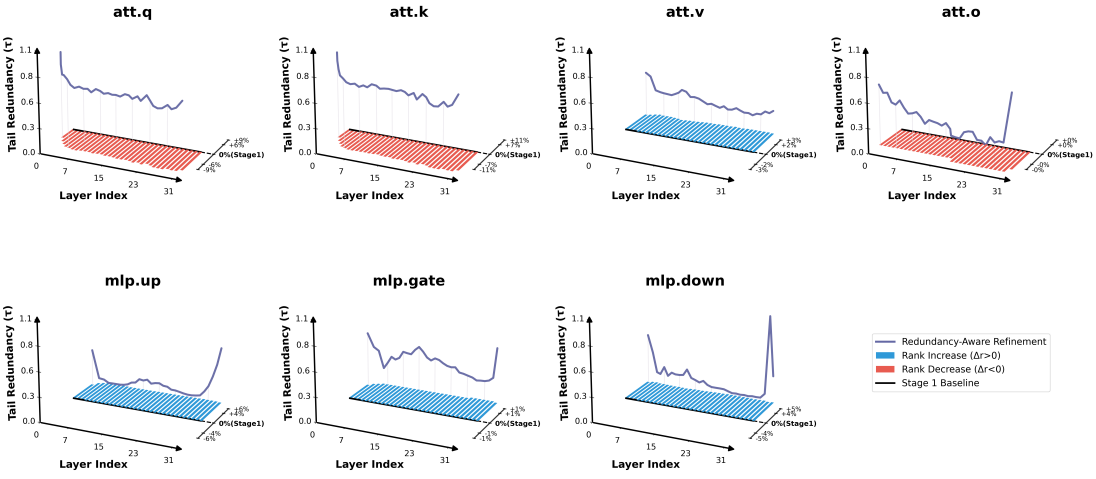

Figure 11: Visualization of Stage II (Redundancy-Aware Refinement). The purple curve shows the Tail Redundancy (τ). The black line represents the allocation baseline established in Stage 1. Deviations are shown as Red bars (rank reduction due to high redundancy) and Blue bars (rank restoration due to low redundancy).