SAES-SVD: SELF-ADAPTIVE SUPPRESSION OF ACCUMULATED AND LOCAL ERRORS FOR SVD-BASED LLM COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid growth in the parameter scale of large language models (LLMs) has created a high demand for efficient compression techniques. As a hardware-agnostic and highly compatible technique, low-rank compression has been widely adopted. However, existing methods typically compress each layer independently by minimizing per-layer reconstruction error, overlooking a critical limitation: the reconstruction error propagates and accumulates through the network, which leads to amplified global deviations from the full-precision baseline. To address this, we propose Self-Adaptive Error Suppression SVD (SAES-SVD), a LLMs compression framework that jointly optimizes intra-layer reconstruction and inter-layer error compensation. SAES-SVD is composed of two novel components: ❶ Cumulative Error-Aware Layer Compression (CEALC), which formulates the compression objective as a combination of local reconstruction and weighted cumulative error compensation. Based on it, we derive a closed-form low-rank solution relied on second-order activation statistics, which explicitly aligns each layer's output with its full-precision counterpart to compensate for accumulated errors. ❷ Adaptive Collaborative Error Suppression (ACES), which automatically adjusts the weighting coefficient to enhance the low-rank structure of the compression objective in CEALC. Specifically, the coefficient is optimized to maximize the ratio between the Frobenius norm of the compressed layer's output and that of the compression objective under a fixed rank, thus ensuring that the rank budget is utilized effectively. Extensive experiments across multiple LLM architectures and tasks show that, without fine-tuning or mixed-rank strategies, SAES-SVD consistently improves post-compression performance. For example, at a 0.2 compression ratio on LLaMA-7B, existing methods exhibit an average accuracy drop exceeding 0.05, whereas SAES-SVD restricts the drop to only 0.02. These improvements underscore the potential of SAES-SVD to effectively narrow the gap between compressed models and their full-precision counterparts, paving the way for more reliable compression of LLMs.

1 INTRODUCTION

Large language models (LLMs), including GPT Family (Achiam et al., 2023; Dettmers et al., 2022), LLaMA Family (Touvron et al., 2023a;b; Dubey et al., 2024), and Qwen Family (Bai et al., 2023; Team, 2024; Yang et al., 2025), have achieved state-of-the-art performance in tasks such as commonsense reasoning, document summarization, and few-shot learning. However, these advances come at a substantial cost: LLMs with billions of parameters demand enormous computational and memory resources, rendering them impractical for deployment on edge devices. To address this challenge, model compression has become a key research direction, with approaches such as quantization (Frantar et al., 2022; Hu et al., 2025), network pruning (Ashkboos et al., 2024), knowledge distillation (Hinton et al., 2014), and low-rank decomposition (Yuan et al., 2023; Li et al., 2025). Among these, low-rank compression, typically implemented through Truncated Singular Value Decomposition (Eckart & Young, 1936), directly reduces both parameter count and computational cost by factorizing weight matrices into low-rank components. This efficiency makes it particularly appealing for LLMs deployment, as the method is hardware-agnostic and can be seamlessly integrated with other compression techniques.

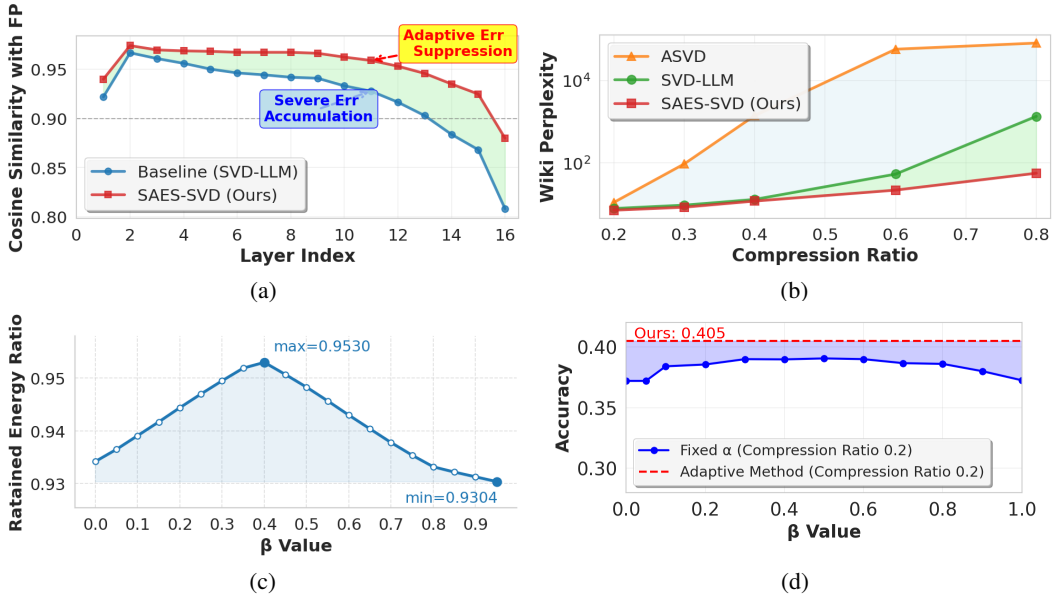

Figure 1: **(a)** Cumulative error analysis on LLaMA-3-8B, showing the cosine similarity of compressed outputs with the original reference across layers. SAES-SVD effectively mitigates error accumulation by adaptively suppressing upstream deviations. **(b)** Perplexity on WikiText2 under varying compression ratios. SAES-SVD consistently achieves lower perplexity than ASVD and SVD-LLM. **(c)** Retained Energy Ratio achieved by using different weighting cofficients for the cumulative error compensation term in with rank=128 compression for k-proj in the 14th Layer of LLaMA-3-8B. Here, for the sake of convenience, we use $\beta = \alpha/(1+\alpha)$ as the horizontal axis. **(d)** Effect of fixed versus adaptive weighting cofficients for the cumulative error compensation on average accuracy across 7 zero-shot benchmarks.

Existing low-rank compression methods typically focus on layer-wise optimization. For example, methods such as ASVD (Yuan et al., 2023), SVD-LLM (Wang et al., 2024), and AdaSVD (Li et al., 2025) aim to reduce local reconstruction error within individual layers by introducing activation-aware scaling, applying whitening transformations, and performing adaptive iterative refinement, respectively. However, these approaches share a fundamental **limitation: they tend to minimize the reconstruction error for each layer independently, without considering how errors propagate and accumulate through the network.** In practice, reconstruction errors from early layers alter the input distribution of subsequent layers, causing these errors to compound and leading to an increasingly larger gap from the full-precision baseline. Applying SVD-LLM (Wang et al., 2024) to LLaMA2-7B further supports this observation. Although SVD-LLM achieves the theoretical minimum truncation error at each layer, the similarity to full-precision outputs, as shown in Figure 1a, decreases steadily from 0.97 to 0.79 across layers. This result indicates that minimizing per-layer reconstruction error alone does not guarantee end-to-end fidelity.

Building on these observations, we propose Self-Adaptive Error Suppression SVD (SAES-SVD), a low-rank compression framework that incorporates cumulative error compensation into the optimization objective. The core idea of SAES-SVD is to let each layer perceive accumulated compression errors from previous layers and adapt its compression strategy based on its sensitivity to those errors. In this way, conventional independent layer-wise optimization is transformed into a globally coordinated error-suppression mechanism. Specifically, SAES-SVD consists of two main components: ❶ **Cumulative Error-Aware Layer Compression (CEALC):** For each layer, we minimize not only the local reconstruction error but also enforce alignment with the corresponding full-precision reference. This alignment helps compensate for errors propagated from upstream compressed layers. The dual objective is formulated as a weighted Frobenius norm of reconstruction and alignment errors. From this objective, we derive a closed-form solution that effectively suppresses reconstruction error and compensates for cumulative error. Moreover, it achieves this with only minimal additional cost, as the method relies solely on second-order statistics of activation deviations and a few extra matrix operations compared with vanilla SVD compression. ❷ **Adaptive Collaborative Error Suppression (ACES):** Since each layer has different sensitivity to compression ratio and cumulative error, using a fixed weighting coefficient (i.e., α_ℓ in Equation 3) for cumulative error compensation often results in suboptimal solutions. As shown in Figure 1c, varying the coefficient across layers leads to different values of the Retained Energy Ratio (RER),

defined as the proportion of the squared sum of the top-k singular values to the total squared sum after SVD. RER serves as a measure of the low-rank structure of the compression objective, where a higher RER indicates stronger energy concentration in the leading singular values and thus greater compression efficiency. To address this issue, we introduce an adaptive weighting coefficient for each layer (i.e., β_ℓ^* in Equation 14). The coefficient is automatically tuned to maximize RER. By concentrating energy on the dominant singular values, this optimization yields compact and informative low-rank representations and enables energy-aware error compensation that improves compression efficiency under fixed rank budgets.

We conducted a comprehensive evaluation of SAES-SVD across diverse datasets, model architectures, and representative SVD-based baselines. The evaluation included language modeling, classification, and generation tasks. Results from multiple benchmarks, such as Figure 1, show that SAES-SVD effectively suppresses cumulative errors introduced during compression. Figure 1a indicates that cosine similarity with full-precision outputs remains consistently higher across layers. Figure 1b and Figure 1d further shows that SAES-SVD achieves much lower perplexity and higher accuracy under diverse compression ratios. The main contributions of this work are threefold:

- We identify that independent layer-wise compression leads to the amplification of cumulative errors. To address this, we propose CEALC, which moves beyond independent optimization and explicitly incorporates cross-layer error accumulation into the low-rank objective. Building on this formulation, we derive an efficient closed-form low-rank solution that depends only on second-order activation statistics.
- We introduce ACES, which use an adaptive coefficient α to control the weighting of cumulative error compensation. The value of α is adjusted to maximize the proportion of retained energy. This optimization enhances the low-rank structure of the compression objective and ensures that, under a fixed rank budget, the retained components preserve as much information as possible.
- CEALC and ACES together form the framework SAES-SVD. Experiments demonstrate that SAES-SVD is both effective and broadly applicable. At the 20% compression ratio, without requiring complex rank allocation or additional fine-tuning, it reduces the drop in zero-shot accuracy by 58% and lowers the perplexity gap by 52% compared with Dip-SVD. These improvements substantially narrow the gap between the compressed model and the full-precision baseline.

2 Related Works

Large Language Model Compression Techniques. The rapid growth of LLMs has spurred extensive research on compression methods that reduce model size and inference cost while preserving performance. Existing approaches can be broadly categorized into four classes: pruning, quantization (Nagel et al., 2021; Frantar et al., 2022; Hu et al., 2025; Lin et al., 2023; Hu et al., 2024), knowledge distillation (Hinton et al., 2014), and low-rank approximation (Yuan et al., 2023; Wang et al., 2024). Unstructured pruning (Ashkboos et al., 2024; Zhang et al., 2023) removes individual weights, but its irregular sparsity patterns often yield limited hardware benefits. Structured pruning (Ma et al., 2023; Dettmers et al., 2023) eliminates entire channels or blocks, improving efficiency but risking significant accuracy loss. Quantization (Frantar et al., 2022; Lin et al., 2023; Hu et al., 2025) reduces numerical precision, offering strong memory savings but typically restricted to 2–8 bits, which limits flexibility and may degrade performance under aggressive settings. Knowledge distillation transfers knowledge from a large teacher model to a smaller student (Liu et al., 2024), but requires costly retraining. In contrast, low-rank approximation, especially singular value decomposition (SVD), provides a hardware-agnostic, post-training alternative that can be combined with other methods, making it a particularly attractive solution for LLM compression.

SVD-based Compression for LLMs. SVD (Golub et al., 1987) reduces matrix sizes by truncating the smallest singular values. Low-rank approximation using SVD has been widely studied as a hardware-agnostic and efficient approach for compressing LLMs. Early works mainly focused on minimizing the per-layer truncation loss. ASVD (Yuan et al., 2023) introduced a learnable scaling matrix to mitigate the loss of discarded singular values, but the improvement remained local and could not guarantee the theoretical minimum truncation loss across all layers. SVD-LLM (Wang et al., 2024; 2025b) addressed this limitation by incorporating a whitening matrix that normalizes input activations, thereby achieving theoretically minimal truncation error. AdaSVD (Li et al., 2025) introduces an alternating optimization scheme that iteratively updates low-rank matrices and com-

pensation terms, thereby progressively reducing compression error. Several subsequent methods have incorporated importance weighting. For example, FW-SVD (Hsu et al., 2022) and GFW-SVD (Chekalina et al., 2025) leverage global importance indicators to guide single-layer compression, whereas DipSVD (Ding et al., 2025) jointly exploits both local and global importance to refine the weighting process. Although these techniques improve layer-level stability, they still rely on the assumption that layer errors are independent, leaving the problem of cumulative error unresolved.

3 PRELIMINARIES

For a weight matrix $W \in \mathbb{R}^{m\times n}$, its SVD is given by $W = U\Sigma V^\top$, where $U \in \mathbb{R}^{m\times m}$ is the left singular vector matrix, $V \in \mathbb{R}^{n\times n}$ is the right singular vector matrix, and $\Sigma \in \mathbb{R}^{m\times n}$ is a nonnegative diagonal matrix with entries arranged in descending order (with $\min(m,n)$ effective diagonal elements). The classical low-rank approximation problem can be formulated as

$$\min_{A\in\mathbb{R}^{m\times r},\, B\in\mathbb{R}^{r\times n}} \|AB - W\|_F^2, \quad r \le \min(m,n). \tag{1}$$

According to the Eckart–Young–Mirsky (Eckart & Young, 1936) theorem, the rank-r optimal approximation of W is $[W]_r = U_r\Sigma_r V_r^\top$. where $U_r \in \mathbb{R}^{m\times r}$ and $V_r \in \mathbb{R}^{n\times r}$ contain the top-r left and right singular vectors of W. By choosing $A = U_r\Sigma_r^{1/2}$ and $B = \Sigma_r^{1/2}V_r^\top$, we obtain $AB = [W]_r$, thus achieving a rank-r decomposition for compressing W. However, direct vanilla SVD compression ignores the statistical variations of layer input activations and often leads to significant accuracy degradation in LLM compression (Yuan et al., 2023; Wang et al., 2024). To address this issue, recent methods formulate LLM compression as an activation-aware, layer-wise independent optimization problem with the following objective:

$$\min_{A_\ell,B_\ell} \|A_\ell B_\ell X_\ell - W_\ell X_\ell\|_F^2, \tag{2}$$

here, X_ℓ denotes the input activations to the ℓ-th layer, generated by the compressed upstream layers. In practice, compression in upstream layers introduces distributional shifts in these inputs. As depth increases, the gap between X_ℓ and its full-precision counterpart X_ℓ^f accumulates, eventually causing a substantial global deviation of the model outputs from the full-precision baseline.

4 METHOD

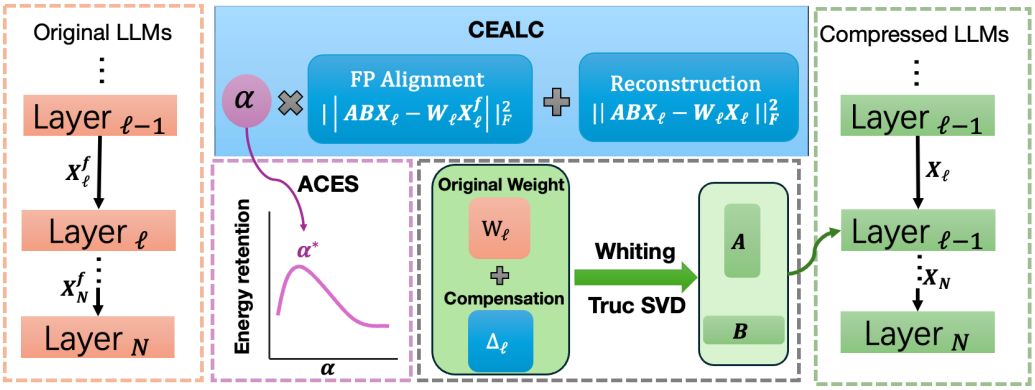

Figure 2: Overview of the proposed SAES-SVD framework.

We propose SAES-SVD, a unified low-rank compression framework that suppresses both per-layer's reconstruction errors and cumulative deviations accumulated from compressed upstream layers. As illustrated in Figure 2, the framework comprises two key components. The first, Cumulative Error-Aware Layer Compression (CEALC), explicitly incorporates cumulative errors compensation into the compression objective, ensuring that each layer aligns closely with its full-precision counterpart. The second, Adaptive Collaborative Error Suppression (ACES), introduces an adaptive coefficient α to balance local reconstruction fidelity with alignment to the full-precision counterpart. The optimal α is chosen to maximize the proportion of ratained energy within the truncated subspace.

4.1 CUMULATIVE ERROR- AWARE LAYER COMPRESSION

Layer compression objective with cumulative error awareness Traditional SVD-based compression methods (Li et al., 2025; Wang et al., 2024) typically approximates the output of the current layer's weight matrix on its input X_ℓ by minimizing Equation 2. However, this approach neglects cumulative errors propagated from upstream compressed layers, which accumulate with increasing depth and amplify the global deviation from the full-precision (FP) baseline. To address this issue, we introduce an additional alignment term with respect to the FP reference output, leading to the following weighted objective:

$$\arg\min_{A_\ell,B_\ell} \underbrace{\|(A_\ell B_\ell - W_\ell)X_\ell\|_F^2}_{\text{intra-layer reconstruction error}} + \alpha_\ell \underbrace{\|A_\ell B_\ell X_\ell - W_\ell X_\ell^f\|_F^2}_{\text{FP reference alignment}}. \tag{3}$$

Here, $\alpha_\ell \geq 0$ controls the alignment strength. Let $T_\ell = W_\ell X_\ell$ be the original output under current inputs and $R_\ell = W_\ell X_\ell^f$ under FP inputs. Then Equation 3 is equivalent to

$$\|(A_\ell B_\ell X_\ell - T_\ell)\|_F^2 + \alpha_\ell \|(A_\ell B_\ell X_\ell - R_\ell)\|_F^2 = (1+\alpha_\ell)\,\|A_\ell B_\ell X_\ell - Z_\ell\|_F^2 + C_\ell, \tag{4}$$

where the constant $C_\ell := \|T_\ell\|_F^2 + \alpha_\ell \|R_\ell\|_F^2 - (1+\alpha_\ell)\|Z_\ell\|_F^2$ is independent of A_ℓ, B_ℓ, and

$$Z_\ell \;=\; \frac{T_\ell + \alpha_\ell R_\ell}{1+\alpha_\ell} \;=\; W_\ell \,\frac{X_\ell + \alpha_\ell X_\ell^f}{1+\alpha_\ell}. \tag{5}$$

Therefore, Equation 3 is equivalent to

$$\min_{A_\ell,B_\ell} \|A_\ell B_\ell X_\ell - Z_\ell\|_F^2. \tag{6}$$

Theorem 4.1. *As illustrated by the detailed derivation in the Appendix A.2, let $H_\ell = X_\ell X_\ell^\top \succ 0$ and $L_\ell = H_\ell^{-1/2}$. Then, for any rank constraint r_ℓ,*

$$\arg\min_{\text{rank}(U_\ell)\leq r_\ell} \|U_\ell X_\ell - Z_\ell\|_F^2 \;=\; \arg\min_{\text{rank}(V_\ell)\leq r_\ell} \left\|V_\ell - M_\ell L_\ell\right\|_F^2, \quad \textit{with } V_\ell := U_\ell H_\ell^{1/2}. \tag{7}$$

Combined with Theorem 4.1, Equation 6 is equivalent to

$$\arg\min_{A,B} \|ABL^{-1} - ZX^\top L\|_F^2 \tag{8}$$

Eliminating explicit inputs via second-order statistics. Storing raw activations X_ℓ and their FP counterparts X_ℓ^f requires prohibitive memory overhead. For instance, even with a moderate calibration set of 128 samples and sequence length 2048, the activation storage cost exceeds the parameter size by more than two orders of magnitude, making it infeasible for practical large-scale compression. To address this, we reformulate the objective in terms of compact second-order statistics. Specifically, we define the input covariance $H_\ell = X_\ell X_\ell^\top$ and the differential covariance $\Delta_\ell = (X_\ell^f - X_\ell)X_\ell^\top$, which capture the distributional characteristics and upstream-induced shifts. By substituting $X_\ell^f X_\ell^\top = H_\ell + \Delta_\ell$, we obtain

$$Z_\ell X_\ell^\top = \frac{W_\ell\big(X_\ell X_\ell^\top + \alpha_\ell X_\ell^f X_\ell^\top\big)}{1+\alpha_\ell} = W_\ell\Big(H_\ell + \beta_\ell \Delta_\ell\Big), \qquad \beta_\ell := \frac{\alpha_\ell}{1+\alpha_\ell} \in [0,1). \tag{9}$$

Substituting Equation 8 into Equation 6, then the optimization objective becomes

$$\arg\min_{A_\ell,B_\ell} \|A_\ell B_\ell H_\ell^{1/2} - W_\ell(H_\ell + \beta\Delta_\ell)H_\ell^{-1/2}\|_F^2 \tag{10}$$

Hence, the target in Equation 6 depends only on H_ℓ and Δ_ℓ, which avoids explicit storage of X_ℓ and X_ℓ^f. In practice, these two second-order statistics are collected in batches and layer by layer, similar to the strategy used in GPTQ, which greatly reduces space complexity.

Closed-form via truncated SVD. Letting $T_\ell = A_\ell B_\ell H_\ell^{1/2}$, and defining the target matrix $G_\ell = W_\ell(H_\ell + \beta\Delta_\ell)H_\ell^{-1/2}$, the Eckart–Young–Mirsky (Eckart & Young, 1936) theorem guarantees that the best rank-r approximation of G_ℓ is obtained by truncated SVD:

$$T_l^* = [G_\ell]_{r_\ell} = \tilde{U}_{r_\ell}\,\Sigma_{r_\ell}\,\tilde{V}_{r_\ell}^\top, \tag{11}$$

where Σ_{r_ℓ} contains the top-r_ℓ singular values, and $\tilde{U}_{r_\ell}, \tilde{V}_{r_\ell}$ are the corresponding left and right singular vectors. Consequently, a closed-form solution for A_ℓ and B_ℓ is given by

$$A_\ell = \tilde{U}_{r_\ell}\,\Sigma_{r_\ell}^{1/2}\,, \qquad B_\ell = \Sigma_{r_\ell}^{1/2}\,\tilde{V}_{r_\ell}^\top\,L_\ell \tag{12}$$

with $L_\ell = H_\ell^{-1/2}$ precomputed from input statistics. At this point, we complete the derivation of the closed-form solution of CEALC. As shown in Equation 10, the approximation target consists of two weighting terms: $W_\ell H_\ell$ and $\beta\Delta_\ell$. The coefficient β controls the strength of compensation for upstream cumulative errors when compressing the current layer ℓ. After applying the whitening matrix $H^{-1/2}$, a truncated SVD yields the optimal theoretical solution.

4.2 Adaptive Collaborative Error Suppression

Motivation. The formulation in Section 4.1 assumes that the alignment coefficient $\beta_\ell = \frac{\alpha_\ell}{1+\alpha_\ell}$ is fixed in advance. In practice, however, different layers have different sensitivities to accumulated errors from upstream layers. A single alignment strength can disrupt the low-rank structure of the original weights. As shown in Figure 1c, varying β alters the low-rank characteristics of the compression objective in Equation 10. An inappropriate choice of β_ℓ scatters spectral energy into non-dominant components, which in turn reduces compression efficiency. To address this, we reinterpret Equation 10 as an approximation problem for the original weight matrix W_ℓ. The key motivation is to permit moderate weight deviations, but only when such deviations help

- ① steers the compressed output toward the FP reference, thereby mitigating cumulative error; and
- ② maximizes the spectral energy preserved in the top-r_ℓ singular subspace, avoiding wasteful energy leakage into discarded components.

CEALC naturally satisfies condition ①, where the coefficient β_ℓ controls the strength of the alignment. Thus, our goal is to determine the optimal β_ℓ such that, under a rank budget r_ℓ, the retained subspace preserves as much information as possible. Formally, we define

$$G_\ell(\beta_\ell) \;=\; S_\ell + \beta_\ell D_\ell, \quad \text{with} \quad \begin{cases} S_\ell = W_\ell H_\ell L_\ell, \\ D_\ell = W_\ell \Delta_\ell L_\ell, \\ \beta_\ell = \dfrac{\alpha_\ell}{1+\alpha_\ell} \in [0,1]. \end{cases} \tag{13}$$

The optimization target can then be expressed as maximizing the proportion of spectral energy captured by the top-r_ℓ singular values of $G_\ell(\beta_\ell)$:

$$\beta_\ell^\star \;=\; \arg\max_{\beta_\ell \in [0,1]} \frac{\sum_{i=1}^{r_\ell} \sigma_i^2\big(G_\ell(\beta_\ell)\big)}{\sum_{i=1}^{\min(m_\ell,n_\ell)} \sigma_i^2\big(G_\ell(\beta_\ell)\big)}, \tag{14}$$

where $\sigma_i(\cdot)$ denotes the i-th singular value in decending order after SVD. **However, singular values change with β, and each trial value would require a new SVD computation, making direct optimization intractable.**

Theorem 4.2 (First-Order Approximation of Fixed Subspace(FS-FOA))**.** *Let $S_\ell \in \mathbb{R}^{m\times n}$ with singular value decomposition $S_\ell = U_\ell \Sigma_\ell V_\ell^\top$, and let $r_\ell \in \{1,\dots,\min(m,n)-1\}$ be the target rank. Denote by $U_{r_\ell} \in \mathbb{R}^{m\times r_\ell}$ and $V_{r_\ell} \in \mathbb{R}^{n\times r_\ell}$ the matrices formed by the top-r_ℓ left and right singular vectors of S_ℓ, respectively. Define the orthogonal projection matrices*

$$P_L^\ell = I - U_{r_\ell}U_{r_\ell}^\top, \qquad P_R^\ell = I - V_{r_\ell}V_{r_\ell}^\top. \tag{15}$$

For $G_\ell(\beta) = S_\ell + \beta D_\ell$, the relative error ratio can be approximated by

$$\widetilde{\rho}_\ell(\beta) \;=\; \frac{\|P_L^\ell\, G_\ell(\beta)\, P_R^\ell\|_F}{\|G_\ell(\beta)\|_F} \;=\; \frac{\|S_\ell^\perp + \beta D_\ell^\perp\|_F}{\|S_\ell + \beta D_\ell\|_F}, \tag{16}$$

where $S_\ell^\perp = P_L^\ell S_\ell P_R^\ell$ and $D_\ell^\perp = P_L^\ell D_\ell P_R^\ell$ are the projections of S_ℓ and D_ℓ onto the orthogonal complements of the top-r_ℓ singular subspaces of S_ℓ. Detailed proof can be referred to Appendix A.3.

For computational tractability, we freeze the principal subspace at $\beta = 0$, leveraging Theorem4.2. The resulting approximation ratio can be written as

$$\widetilde{\rho}(\beta) = \frac{\|S_\ell^\perp + \beta D_\ell^\perp\|_F^2}{\|S_\ell + \beta D_\ell\|_F^2} = \frac{a + 2b\,\beta + c\,\beta^2}{A + 2B\,\beta + C\,\beta^2}\,, \tag{17}$$

where

$$a = \|S_{\ell\perp}\|_F^2\,, \quad b = \langle S_{\ell\perp}, D_{\ell\perp}\rangle\,, \quad c = \|D_{\ell\perp}\|_F^2\,, \quad A = \|S_\ell|_F^2, \quad B = \langle S_\ell, D_\ell\rangle, \quad C = \|D_\ell\|_F^2 \tag{18}$$

Setting the derivative of $\widetilde{\rho}(\beta)$ to zero yields a quadratic equation

$$\widetilde{p}(\beta) = (cB - bC)\,\beta^2 + (cA - aC)\,\beta + (bA - aB) = 0\,. \tag{19}$$

We evaluate all real roots within a feasible interval $[\beta\min, \beta_{\max}]$ (e.g., $[0, 1]$), along with the endpoints and the minimizer $\beta^\star$ of $\widetilde{\rho}(\beta)$ is then selected. This adaptive selection minimizes the fraction of truncated energy, ensuring that the retained singular spectrum captures the maximum proportion of information. As illustrated in Figure 1d, the adaptive strategy consistently outperforms fixed-α settings across different compression ratios, highlighting both its robustness and effectiveness.

Through CEALC and ACES, we establish a closed-loop compression mechanism: each layer adaptively selects α_ℓ to suppress newly introduced truncation error, while residual error is accumulated into second-order statistics and propagated forward. This allows subsequent layers to anticipate and compensate for upstream cumulative errors. Consequently, the model achieves output fidelity close to the baseline using only a single SVD approximation per layer. The detailed implementation can be found in Appendix A.6.

5 Experiments.

Models and Datasets. To comprehensively evaluate the performance of SAES-SVD across diverse models and tasks, we conducted systematic experiments on mainstream open-source LLMs and standard language understanding and generation benchmarks, covering a wide range of parameter budgets and hardware settings. Unless otherwise specified, all results are obtained using a single-pass SVD per layer without task-specific fine-tuning after compression. For model selection, we considered representative architectures including LLaMA-1 (Touvron et al., 2023a), LLama-2 (Touvron et al., 2023b) and LLaMA-3 (LLaMA Team, AI @ Meta, 2024). Regarding evaluation tasks and datasets, we focused on natural language reasoning and understanding benchmarks. Specifically, we measured perplexity on the WikiText2 (Merity et al., 2016) and C4 (Raffel et al., 2020) dataset with a sequence length of 2048 to assess the language coherence and modeling capacity of the compressed models. In addition, we evaluated zero-shot accuracy on multiple datasets, including ARC-Challenge (Clark et al., 2018), ARC-Easy (Clark et al., 2018), HellaSwag (Zellers et al., 2019), MathQA (Amini et al., 2019), PIQA (Bisk et al., 2020), and WinoGrande (Sakaguchi et al., 2021), to comprehensively examine the generalizability of the compressed models on various tasks.

Baselines. In a unified calibration dataset and evaluation protocol, we compared our method with a range of well-established SVD-based low-rank compression approaches, including ASVD (Li et al., 2025), SVD-LLM (Wang et al., 2024), FW-SVD (Hsu et al., 2022), Dobi-SVD (Wang et al., 2025a), and AdaSVD (Li et al., 2025). Notably, SVD-LLM and Dobi-SVD rely on fine-tuning, while A-SVD, AdaSVD and Dobi-SVD employ mixed-rank strategies. In contrast, our proposed SAES-SVD does not require additional training nor auxiliary techniques, yet it consistently outperforms these baselines, highlighting its effectiveness and simplicity.

Main Results. We conducted a comprehensive evaluation of SAES-SVD for low-rank compression of LLMs and compared it with several SVD-based baselines across different compression ratios. Overall, SAES-SVD consistently achieved superior performance across tasks and model scales. Table 1 summarizes results on the representative LLaMA-7B. SAES-SVD reduced the perplexity gap by more than 35% compared with the strongest baseline and lowered the accuracy drop by 30–40%. At the more challenging 0.6 ratio, it halved the perplexity gap relative to Dobi-SVD and SVD-LLM, while keeping the accuracy drop below 0.03, compared with more than 0.05 for other methods. These results lead to two key observations: (i) SAES-SVD consistently reduces perplexity by a large margin—up to a 2× improvement at aggressive compression levels—and (ii) it preserves or

Table 1: Perplexity and zero-shot evaluation of LLaMA-7B across seven benchmark datasets under varying compression ratios. The table compares SAES-SVD with competing SVD-based methods (ASVD* (Yuan et al., 2023), FWSVD (Hsu et al., 2022), SVD-LLM† (Wang et al., 2024), Dobi-SVD*† (Wang et al., 2025a), Dip-SVD*). Methods with fine-tuning are marked by †; mixed-rank strategies by *. Due to the closed-source implementation of Dip-SVD, we were unable to obtain its performance at the 0.4 compression ratio.

Ratio	Method	Wiki2↓	PTB↓	C4↓	Openb.↑	ARC_e↑	ARC_c↑	WinoG.↑	HellaS.↑	PIQA↑	MathQA↑	Avg.↑	Drop↓
0.0	Baseline	5.68	8.35	7.34	0.28	0.67	0.38	0.67	0.56	0.78	0.27	0.52	0.00
0.2	FWSVD	2e5	3e4	2e3	0.09	0.11	0.06	0.05	0.08	0.10	0.05	0.02	96%
	ASVD†	11.14	16.55	15.93	0.25	0.53	0.27	0.64	0.41	0.68	0.24	0.43	16.7%
	SVD-LLM†	7.94	16.22	15.84	0.22	0.58	0.29	0.63	0.43	0.69	0.24	0.44	14.7%
	Dobi-SVD*†	8.54	14.83	10.01	0.26	0.59	0.31	**0.66**	0.44	0.70	0.23	0.46	10.8%
	Dip-SVD*	7.95	15.60	14.07	0.27	0.63	0.33	0.64	0.45	0.71	0.24	0.47	9.2%
	Ours	**7.17**	**15.16**	**13.77**	**0.29**	**0.68**	**0.36**	0.65	**0.45**	**0.75**	**0.25**	**0.50**	**3.9%**
0.4	FWSVD	2e4	1e4	1e4	0.06	0.05	0.02	0.02	0.00	0.05	0.03	0.02	96.6%
	ASVD†	1e3	3e3	1e3	0.13	0.28	0.22	0.48	0.26	0.55	0.19	0.30	41.8%
	SVD-LLM†	13.11	63.75	49.83	0.19	0.42	0.25	0.58	0.33	0.60	0.21	0.37	28.3%
	Dobi-SVD*†	13.54	46.38	23.54	0.22	0.41	0.27	0.58	0.34	0.61	0.23	0.38	26.3%
	Dip-SVD*	12.76	46.95	34.35	0.22	0.50	**0.30**	0.61	0.36	0.64	0.22	0.40	22.8%
	Ours	**10.42**	**45.13**	**32.79**	**0.23**	**0.50**	0.29	**0.62**	**0.36**	**0.65**	**0.23**	**0.41**	**21.1%**
0.6	FWSVD	3e4	2e4	2e4	0.06	0.01	0.00	0.00	0.01	0.01	0.00	0.01	97.8%
	ASVD†	6e4	4e4	4e5	0.12	0.26	0.21	0.49	0.26	0.53	0.18	0.29	43.8%
	SVD-LLM†	53.74	4e2	3e2	0.14	0.28	0.22	0.50	0.27	0.55	0.21	0.31	39.9%
	Dobi-SVD*†	46.18	2e2	2e2	0.15	0.31	0.20	0.52	0.28	0.54	0.22	0.32	38.0%
	Ours	**22.01**	**116.83**	**93.97**	**0.16**	**0.33**	**0.25**	**0.52**	**0.30**	**0.54**	**0.23**	**0.34**	**34.1%**

Table 2: Zero-shot accuracy of LLaMA-13B under different compression ratios on seven benchmark datasets. Results are reported for SVD-LLM†, Dip-SVD*, and our proposed method. † denotes methods that rely on fine-tuning, while * indicates methods using mixed-rank strategies.

Ratio	Method	Openb.	ARC_e	WinoG.	HellaS.	ARC_c	PIQA	MathQA	Average
0.2	SVD-LLM†	0.302	0.683	0.684	0.470	0.356	0.725	0.265	0.498
	Dip-SVD*	0.306	0.681	**0.692**	**0.490**	0.369	0.734	0.258	0.504
	Ours	**0.308**	**0.712**	0.684	0.477	**0.388**	**0.734**	**0.265**	**0.510**
0.4	SVD-LLM†	0.222	0.521	0.639	0.355	0.248	0.637	0.228	0.407
	Dip-SVD*	0.230	0.548	0.644	0.402	0.283	**0.661**	0.233	0.429
	Ours	**0.248**	**0.543**	**0.654**	**0.470**	**0.284**	0.631	**0.239**	**0.438**

even improves accuracy, while other methods suffer notable degradation. Crucially, SAES-SVD outperforms not only methods requiring fine-tuning (e.g., SVD-LLM, Dobi-SVD) but also those relying on mixed-rank allocations (e.g., AdaSVD, Dobi-SVD,Dip-SVD). These findings, together with additional experiments shown in Table 5 and Table 6 in appendix, confirm that the gains of SAES-SVD generalize across architectures and model scales, underscoring both its robustness and broad applicability.

Comparison on larger-scale models. We further evaluated our method on larger architectures, including LLaMA-13B and LLaMA-30B, against current SOTA approaches. As shown in Table 2 and Figure 3, we conducted experiments across compression ratios ranging from 0.2 to 0.4. Our method consistently outperforms all baselines in terms of both average accuracy and perplexity. Notably, on LLaMA-30B, SAES-SVD reduces perplexity to 5.49, compared with 5.63 for SVD-LLM and 6.64 for Dip-SVD, while achieving a 10% accuracy improvement over Dip-SVD. These results demonstrate that SAES-SVD remains highly effective at larger scales and delivers substantial gains in compressed model performance.

Inference speed. Low-rank approximation simultaneously decreases both computational complexity and parameter storage. This dual reduction substantially lowers memory footprint and compute requirements. We evaluated the inference speed of LLaMA3-8B under different compression ratios on a single NVIDIA A6000 GPU. As illustrated in Figure 4, our proposed SAES-SVD achieves consistent acceleration over the FP16 baseline, with speedups ranging from 1.29× to 3.79× as the compression ratio increases. These results demonstrate the practicality of SAES-SVD in accelerating inference while significantly reducing memory demands.

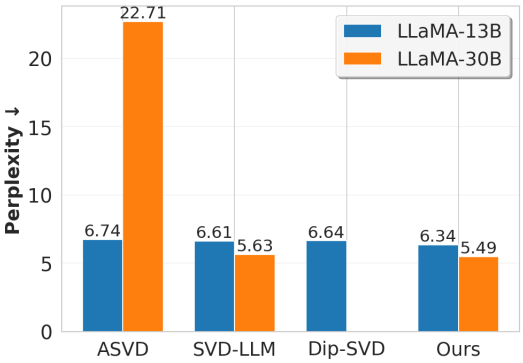

Figure 3: Wiki-PPL for two large-scale models, across different methods at 20% compression.

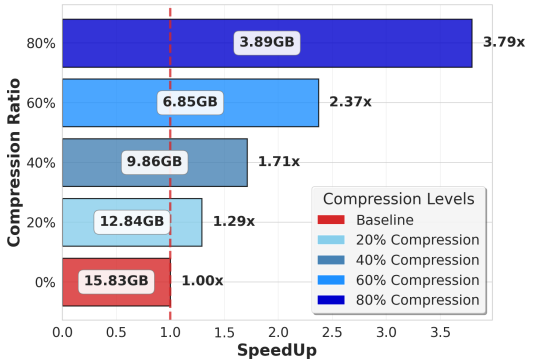

Figure 4: Memory usage and inference speedup of LLaMA-3-8B on varying compression ratios.

Table 3: Ablation study on the two components of our method. Avg Acc denotes the average accuracy over PIQA, ARC-e, HellaSwag and WinoGrande.

Model	CEALC	ACES	Wiki PPL↓	Avg Acc↑
LLama2-7B	×	×	9.34	58.66
	✓	×	7.66	62.02
	✓	✓	7.37	63.03
LLama3-8B	×	×	16.59	55.76
	✓	×	12.25	58.82
	✓	✓	11.48	60.18

Table 4: Sensitivity analysis of the upper and lower bounds of the adaptive coefficient α in ACES, conducted on LLaMA2-7B at a 20% compression ratio.

$\alpha_{\min}$	$\alpha_{\max}$	Wiki PPL↓	Avg Acc↑
0	1	8.96	58.47
0.1	0.9	7.88	62.81
0.2	0.8	7.38	63.02
0.25	0.75	7.37	63.03
0.3	0.7	7.36	63.01
0.4	0.6	7.63	62.87

Ablation Experiments. The proposed SAES-SVD framework consists of two complementary components: Cumulative Error-Aware Layer Compression (CEALC) and Adaptive Collaborative Error Suppression (ACES). CEALC addresses the limitation of prior methods that treat layer compression as independent processes; as shown in Figure 1a, it effectively mitigates cumulative compression errors. Building on this, ACES further enhances the low-rank properties of the compression objective shaped by CEALC, leading to superior performance under limited rank budgets, as illustrated in Figure 1d. The ablation results presented in Table 3 confirm these contributions: CEALC significantly improves both generative and zero-shot tasks, while ACES provides additional performance gains on top of CEALC.

We conducted a sensitivity analysis of the adaptive coefficient α by varying its lower and upper bounds. Experiments were performed on LLaMA-2 7B at a 20% compression ratio, with results summarized in Table 4. The analysis shows that performance is relatively stable within moderate ranges of α. Specifically, when $\alpha_{\min} = 0.25$ and $\alpha_{\max} = 0.75$, the model achieves the best trade-off, with an average zero-shot accuracy of 63.03%. In contrast, setting the lower bound too high (e.g., $\alpha_{\min} > 0.8$) slightly degrades perplexity while reducing average accuracy significantly, indicating potential overfitting calibration dataset. These results validate our choice of setting the lower bound to 0.25 and the upper bound to 0.75 in practice.

6 Conclusion

In this work, we proposed SAES-SVD, a principled framework for low-rank compression of large language models that explicitly addresses both local reconstruction error and cross-layer accumulated error. By introducing Cumulative Error-Aware Layer Compression (CEALC) and Adaptive Collaborative Error Suppression (ACES), our method effectively mitigates error propagation while preserving critical low-rank structure. Extensive experiments across diverse models and tasks demonstrate that SAES-SVD consistently outperforms prior SVD-based methods in terms of perplexity and zero-shot accuracy, even under aggressive compression ratios, and does so without requiring fine-tuning or mixed-rank heuristics. These findings highlight the robustness, scalability, and practicality of SAES-SVD, providing a promising direction for deploying large language models efficiently in real-world scenarios.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi Kohlmeier. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2357–2367, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1227. URL https://aclanthology.org/P19-1227/.

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 7432–7439, 2020.

Viktoriia Chekalina, Daniil Moskovskiy, Daria Cherniuk, Maxim Kurkin, Andrey Kuznetsov, and Evgeny Frolov. Generalized fisher-weighted svd: Scalable kronecker-factored fisher approximation for compressing large language models. *arXiv preprint arXiv:2505.17974*, May 2025. URL https://arxiv.org/abs/2505.17974.

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

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35: 30318–30332, 2022.

Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. Spqr: A sparse-quantized representation for near-lossless llm weight compression. *arXiv preprint arXiv:2306.03078*, 2023.

Xuan Ding, Rui Sun, Yunjian Zhang, Xiu Yan, Yueqi Zhou, Kaihao Huang, Suzhong Fu, Chuanlong Xie, and Yao Zhu. Dipsvd: Dual-importance protected svd for efficient llm compression. *arXiv preprint arXiv:2506.20353*, June 2025. URL https://arxiv.org/abs/2506.20353.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. doi: 10.1007/BF02288367.

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

G. H. Golub, Alan Hoffman, and G. W. Stewart. A generalization of the eckart–young–mirsky matrix approximation theorem. *Linear Algebra and its Applications*, 88/89:317–327, 1987. doi: 10.1016/0024-3795(87)90114-5.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *Advances in Neural Information Processing Systems*, 2014.

Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. Language model compression with weighted low-rank factorization. In *International Conference on Learning Representations (ICLR) 2022 Workshop*, 2022. URL https://arxiv.org/abs/2207.00112. arXiv preprint arXiv:2207.00112.

Xing Hu, Yuan Cheng, Dawei Yang, Zhihang Yuan, Jiangyong Yu, Chen Xu, and Sifan Zhou. I-llm: Efficient integer-only inference for fully-quantized low-bit large language models. *arXiv preprint arXiv:2405.17849*, May 2024. URL https://arxiv.org/abs/2405.17849.

Xing Hu, Yuan Cheng, Dawei Yang, Zukang Xu, Zhihang Yuan, Jiangyong Yu, Chen Xu, Zhe Jiang, and Sifan Zhou. Ostquant: Refining large language model quantization with orthogonal and scaling transformations for better distribution fitting. *arXiv preprint arXiv:2501.13987*, 2025.

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

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. Awq: Activation-aware weight quantization for llm compression and acceleration. *arXiv preprint arXiv:2306.00978*, 2023.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. LLM-QAT: Data-free quantization aware training for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 467–484, Bangkok, Thailand, aug 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.26. URL https://aclanthology.org/2024.findings-acl.26/.

LLaMA Team, AI @ Meta. The llama-3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. In *Advances in Neural Information Processing Systems*, 2023.

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

Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization. *arXiv preprint arXiv:2106.08295*, 2021. doi: 10.48550/arXiv.2106.08295. URL https://arxiv.org/abs/2106.08295.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 2021.

Qwen Team. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

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

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

Xin Wang, Samiul Alam, Zhongwei Wan, Hui Shen, and Mi Zhang. Svd-llm v2: Optimizing singular value truncation for large language model compression. *arXiv preprint arXiv:2503.12340*, March 2025b. URL https://arxiv.org/abs/2503.12340.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Mingyang Zhang, Chunhua Shen, Zhen Yang, Linlin Ou, Xinyi Yu, Bohan Zhuang, et al. Pruning meets low-rank parameter-efficient fine-tuning. *arXiv preprint arXiv:2305.18403*, 2023.

A APPENDIX

A.1 LLM DISCLAIMER

The authors hereby declare the role of large language model (LLM) tools in the preparation of this manuscript: LLMs were solely utilized to assist with text polishing (including refining sentence structure, optimizing lexical expression, and enhancing language fluency) and **writing optimization** of the paper's narrative content.

It is explicitly emphasized that all core components of this research, which determine the originality, scientific validity, and academic value of the work, were independently completed by the research team through manual efforts. These components include, but are not limited to:

- The formulation and development of the overall research framework, core ideas, and logical structure of the study;
- The design, coding, debugging, and validation of all algorithms and program codes involved in the research;
- The design of experimental protocols, collection and preprocessing of experimental data, execution of experiments, analysis and interpretation of experimental results, and verification of conclusions.

The use of LLM tools did not involve any participation in the conception of research content, generation of technical solutions, implementation of experimental processes, or derivation of research conclusions. All content of this paper adheres to academic integrity standards, and the research team assumes full responsibility for the scientificity, authenticity, and originality of the work.

A.2 THEORETICAL PROOF OF THEOREM 4.1 AND SUBSEQUENT TRUNCATED SVD

Problem Formulation. Consider the optimization problem

$$\min_{A,B} J(A,B) = \|ABX - Z\|_F^2, \tag{20}$$

where the goal is to simplify the objective function $J(A,B)$ so that optimization with respect to A and B becomes more tractable. The key idea is to exploit the structure of the matrix X, particularly its row space.

Assumption. To ensure that the following projection operations are well-defined, we assume that X has full row rank, because in practice a small ridge parameter λ will be introduced and $L_\ell = (H_\ell + \lambda I)^{-1/2}$. This assumption is reasonable since the activation matrix X typically has far more columns than rows. If X is a $p \times n$ matrix with $p \ll n$, then $\mathrm{rank}(X) = p$, which guarantees that $XX^\top$ is invertible.

Orthogonal Projection Operator. We define the projection matrix

$$P = X^\top (XX^\top)^{-1} X, \tag{21}$$

which is an $n \times n$ orthogonal projector mapping any row vector in $\mathbb{R}^n$ onto the row space of X. For any matrix Y with n columns, the product YP projects each row of Y onto the row space of X.

The matrix P satisfies several important properties:

- **Symmetry:** $P^\top = P$, since $XX^\top$ is symmetric and so is its inverse.
- **Idempotence:** $P^2 = P$, which follows by direct calculation.
- **Projection property:** For any matrix of the form CX, right multiplication by P leaves it unchanged, i.e., $(CX)P = CX$. In particular, ABX belongs to this class.

Decomposition of Z. We decompose Z into components lying in the row space of X and its orthogonal complement:

$$Z_P = ZP, \qquad Z_\perp = Z(I - P). \tag{22}$$

Clearly,

$$Z = ZP + Z(I - P) = Z_P + Z_\perp, \tag{23}$$

where Z_P is the projection of Z onto the row space of X, and $Z_\perp$ is the projection onto its orthogonal complement. Moreover, $Z_\perp P = 0$.

Reformulated Objective. Using this decomposition, we expand the objective:

$$\begin{aligned} J(A,B) &= \|ABX - Z\|_F^2 \\ &= \|ABX - (Z_P + Z_\perp)\|_F^2 \\ &= \|(ABX - Z_P) - Z_\perp\|_F^2. \end{aligned} \tag{24}$$

Let $U = ABX - Z_P$ and $V = Z_\perp$. Since U lies in the row space of X and V lies in its orthogonal complement, they are orthogonal, i.e., $\mathrm{Tr}(U^\top V) = 0$. By the Pythagorean theorem for the Frobenius norm, we obtain

$$J(A,B) = \|ABX - Z_P\|_F^2 + \|Z_\perp\|_F^2. \tag{25}$$

By the premise, we start from

$$J(A,B) \;=\; \| \, ABX - ZX^\top (XX^\top)^{-1} X \, \|_F^2 \;=\; \| \, UX - ZP \, \|_F^2, \qquad where\, U = AB. \tag{26}$$

Step 1: Projection reduction is tight up to a constant. Decompose Z into its components on the row space of X and its orthogonal complement:

$$Z \;=\; ZP \;+\; Z(I-P) \;=:\; Z_P + Z_\perp, \qquad Z_\perp P = 0. \tag{27}$$

Since UX lies in the row space of X, we have $(UX)P = UX$. Therefore,

$$\begin{aligned}\|UX - Z\|_F^2 &= \|UX - (Z_P + Z_\perp)\|_F^2 = \|(UX - Z_P) - Z_\perp\|_F^2 \\ &= \|UX - Z_P\|_F^2 + \|Z_\perp\|_F^2 \quad \text{(orthogonality in the row/orthogonal subspaces).}\end{aligned} \tag{28}$$

Thus minimizing $\|UX - ZP\|_F^2$ is equivalent to minimizing $\|UX - Z\|_F^2$, since they differ by the constant $\|Z_\perp\|_F^2$ independent of U. Hence equation 26 is a valid surrogate objective.

Step 2: Whitening reformulation. Let $M = ZX^\top \in \mathbb{R}^{m\times p}$. Expand the Frobenius norm using trace identities:

$$\begin{aligned}\|UX - ZP\|_F^2 = \|UX - ZP\|_F^2 = \|(UX - Z)P\|_F^2 &= \operatorname{tr}\big((UX - Z)P(UX - Z)^\top\big) \\ &= \operatorname{tr}(UXX^\top U^\top) - 2\,\operatorname{tr}(UXPZ^\top) + \operatorname{tr}(ZPZ^\top) \\ &= \operatorname{tr}(UHU^\top) - 2\,\operatorname{tr}(UM^\top) + \operatorname{tr}(ZPZ^\top).\end{aligned} \tag{29}$$

Because $H \succ 0$, its symmetric square roots $H^{\pm 1/2}$ exist and are invertible. Using

$$\operatorname{tr}(UHU^\top) = \|UH^{1/2}\|_F^2, \qquad \operatorname{tr}(UM^\top) = \langle UH^{1/2},\, MH^{-1/2}\rangle_F, \tag{30}$$

we can complete the square:

$$\begin{aligned}\|UX - ZP\|_F^2 &= \|UH^{1/2}\|_F^2 - 2\langle UH^{1/2}, MH^{-1/2}\rangle_F + \operatorname{tr}(ZPZ^\top) \\ &= \|UH^{1/2} - MH^{-1/2}\|_F^2 \;+\; \Big(\operatorname{tr}(ZPZ^\top) - \|MH^{-1/2}\|_F^2\Big).\end{aligned} \tag{31}$$

The parenthetical term is independent of U. Therefore the minimizers of equation 26 are exactly the minimizers of

$$\min_{\operatorname{rank}(U)\le r} \big\|UH^{1/2} - MH^{-1/2}\big\|_F^2. \tag{32}$$

At this point, we have completed the proof of Theorem 4.1. Subsequently, we will perform low-rank approximation based on this.

Step 3: Invertible change of variables and rank preservation. Define the change of variables

$$V \;=\; UH^{1/2}, \qquad L \;=\; H^{-1/2}. \tag{33}$$

Since $H^{1/2}$ is invertible, $\operatorname{rank}(V) = \operatorname{rank}(U)$, so the rank constraint is preserved. The problem is thus equivalent to

$$\min_{\operatorname{rank}(V)\le r} \|\,V - ML\,\|_F^2. \tag{34}$$

This proves the desired whitening equivalence: minimizing $\|UX - ZP\|_F^2$ over rank-r U is equivalent to minimizing $\|V - ML\|_F^2$ over rank-r V, with the bijection between minimizers given by $U^\star = V^\star H^{-1/2}$.

A.3 THEORETICAL PROOF OF THEOREM 4.2: FIXED-SUBSPACE FIRST-ORDER APPROXIMATION FOR TRUNCATION RATIOS

Setup. Let $S \in \mathbb{R}^{m\times n}$ with singular value decomposition $S = U\Sigma V^\top$, and fix a target rank $r \in \{1, \dots, \min(m,n)-1\}$. Denote

$$U_r := U_{[:,1:r]}, \quad V_r := V_{[:,1:r]}, \qquad P_L := I - U_r U_r^\top, \quad P_R := I - V_r V_r^\top.$$

Given a direction $D \in \mathbb{R}^{m\times n}$ and a scalar $\beta \in [0,1]$, define the affine path

$$G(\beta) := S + \beta D. \tag{35}$$

Let $\|\cdot\|_F$ be the Frobenius norm, $\|\cdot\|_2$ the spectral norm, and let $\sigma_1(\cdot) \geq \cdots$ denote singular values.

True tail energy and ratio. By the Eckart–Young–Mirsky theorem,

$$E(\beta) \;:=\; \min_{\mathrm{rank}(Q)\leq r} \|G(\beta) - Q\|_F^2 \;=\; \sum_{i>r} \sigma_i\big(G(\beta)\big)^2. \tag{36}$$

We call $\sqrt{E(\beta)}$ the *tail (Frobenius) norm* of $G(\beta)$ at rank r. The corresponding *tail ratio* is

$$\rho(\beta) \;:=\; \frac{\sqrt{E(\beta)}}{\|G(\beta)\|_F}. \tag{37}$$

Fixed-subspace first-order approximation (FS-FOA). Freeze the leading rank-r subspaces at $\beta = 0$ (those of S), and approximate the tail by the projection onto the fixed orthogonal complement:

$$\widetilde{E}(\beta) \;:=\; \big\| P_L\, G(\beta)\, P_R \big\|_F^2 \;=\; \big\| S_\perp + \beta D_\perp \big\|_F^2, \quad S_\perp := P_L S P_R,\;\; D_\perp := P_L D P_R. \tag{38}$$

The *FS-FOA tail ratio* is then

$$\widetilde{\rho}(\beta) \;:=\; \frac{\|S_\perp + \beta D_\perp\|_F}{\|S + \beta D\|_F}. \tag{39}$$

Assumption (spectral gap and small perturbation). Let the spectral gap at r be

$$\delta \;:=\; \sigma_r(S) - \sigma_{r+1}(S) \;>\; 0, \tag{40}$$

and assume the relative perturbation is small:

$$\tau \;:=\; \frac{\beta\,\|D\|_2}{\delta} \;\leq \tau_0, \qquad \text{for some sufficiently small } \tau_0 \in (0,1). \tag{41}$$

Theorem (FS-FOA is first-order accurate for the tail energy). Under equation 40–equation 41, there exist absolute constants $C_1, C_2 > 0$ such that for all $\beta \in [0,1]$,

$$E(\beta) \;=\; \|S_\perp + \beta D_\perp\|_F^2 \;+\; R_E(\beta), \qquad \big|R_E(\beta)\big| \;\leq\; C_1\,\tau^2\Big(\|S\|_F^2 + \beta^2\|D\|_F^2\Big). \tag{42}$$

In particular, the discrepancy between the true tail energy and its FS-FOA surrogate is *second order* in τ.

Corollary (FS-FOA is first-order accurate for the tail ratio). With the same assumptions, there exists $C' > 0$ such that

$$\big|\rho(\beta) - \widetilde{\rho}(\beta)\big| \;\leq\; C'\,\tau^2 \;=\; O\!\left(\frac{\beta^2\|D\|_2^2}{\delta^2}\right). \tag{43}$$

Hence $\widetilde{\rho}(\beta)$ is a first-order accurate approximation to the true ratio $\rho(\beta)$.

Proof sketch of equation 42. Let $U_r(\beta), V_r(\beta)$ be the leading r-dimensional singular subspaces of $G(\beta)$ and $P_L(\beta) = I - U_r(\beta)U_r(\beta)^\top$, $P_R(\beta) = I - V_r(\beta)V_r(\beta)^\top$ their orthogonal complements. Standard Davis–Kahan/Wedin perturbation bounds imply

$$\|P_L(\beta) - P_L\|_2 \lesssim \tau, \qquad \|P_R(\beta) - P_R\|_2 \lesssim \tau. \tag{44}$$

Because $E(\beta) = \|P_L(\beta)G(\beta)P_R(\beta)\|_F^2$, write $P_L(\beta) = P_L + E_L$, $P_R(\beta) = P_R + E_R$ with $\|E_{L/R}\|_2 = O(\tau)$ and expand:

$$P_L(\beta)G(\beta)P_R(\beta) = (P_L + E_L)(S + \beta D)(P_R + E_R) = P_L(S + \beta D)P_R + \mathcal{E},$$

where the Frobenius norm of the remainder satisfies $\|\mathcal{E}\|_F = O\big(\tau \|G(\beta)\|_F\big)$ by the block structure induced by (U_r, V_r). Squaring norms yields

$$E(\beta) = \|P_L G P_R\|_F^2 + O\big(\tau^2 \|G(\beta)\|_F^2\big) = \|S_\perp + \beta D_\perp\|_F^2 + O\big(\tau^2(\|S\|_F^2 + \beta^2\|D\|_F^2)\big),$$

which is equation 42. The corollary equation 43 follows by dividing by $\|G(\beta)\|_F^2$ and using $\sqrt{1+x} = 1 + O(x)$ for small x.

Closed form for the FS-FOA ratio and its stationary condition. Both numerator and denominator of equation 39 are quadratic polynomials in β:

$$\|S_\perp + \beta D_\perp\|_F^2 = a + 2b\,\beta + c\,\beta^2, \tag{45}$$

$$\|S + \beta D\|_F^2 = A + 2B\,\beta + C\,\beta^2, \tag{46}$$

where

$$a := \|S_\perp\|_F^2, \quad b := \langle S_\perp, D_\perp\rangle, \quad c := \|D_\perp\|_F^2, \qquad A := \|S\|_F^2, \quad B := \langle S, D\rangle, \quad C := \|D\|_F^2. \tag{47}$$

Let

$$\phi(\beta) := \widetilde{\rho}(\beta) = \sqrt{\frac{a + 2b\beta + c\beta^2}{A + 2B\beta + C\beta^2}}. \tag{48}$$

Since the square root is monotone, $\arg\min \phi$ coincides with $\arg\min$ of the squared ratio. Differentiating and simplifying gives the *quadratic* stationary equation

$$(cB - bC)\,\beta^2 + (cA - aC)\,\beta + (bA - aB) = 0. \tag{49}$$

Selection rule. Solve equation 49, keep real roots in $[0, 1]$, add the boundaries $\{0, 1\}$, evaluate $\widetilde{\rho}(\beta)$ on this candidate set, and choose the minimizer $\beta^* \in [0, 1]$. In applications with a trade-off parameter $\alpha \geq 0$, one uses the bijection

$$\beta = \frac{\alpha}{1+\alpha} \quad \Longleftrightarrow \quad \alpha = \frac{1}{1-\beta} - 1, \tag{50}$$

so that $\alpha^* = \frac{1}{\beta^*} - 1$.

When is FS-FOA accurate. Accuracy is controlled by $\tau = \beta\|D\|_2/\delta$. When $\tau \ll 1$ (e.g., $\tau \lesssim 0.5$), the error in equation 42–equation 43 is second order and the stationary points of $\widetilde{\rho}$ reliably approximate those of ρ. In near-degenerate cases (small spectral gap or large D), one may constrain $\beta \leq \beta_{\max} < 1$ or apply a mild shrinkage to the selected β^*; a one-step refinement that recomputes the leading subspaces at $G(\beta^*)$ further reduces the residual error.

Interpretation. The numerator in equation 39 measures the energy that would be *discarded* by rank-r truncation (tail) when measured in the fixed orthogonal complement of the leading subspaces of S; the denominator measures the *total* energy. The quadratic form equation 49 encodes the balance between tail alignment (a, b, c) and total alignment (A, B, C), yielding a closed-form, one-dimensional selection of β (hence α) that minimizes the truncation ratio to first order.

A.4 QUANTITATIVE RESULTS

Table 5: Performance of Compressed Llama-2-7B Across Benchmarks and Compression Ratios.

Ratio	Method	WikiText-2	PTB	C4	ARC_e	WinoG.	HellaS.	PIQA	Average
0	Original	5.68	8.35	7.34	74.62	69.22	76.00	79.11	68.85
0.4	SVD	39661.03	69493.00	56954.00	26.39	48.62	25.64	52.99	38.41
	FWSVD	8060.35	9684.10	7955.21	26.05	50.20	25.70	52.39	38.59
	ASVD	1609.32	7319.49	1271.85	26.81	49.49	25.83	53.81	38.99
	SVD-LLM	161.11	719.44	61.95	36.99	56.04	30.49	56.96	45.12
	AdaSVD	14.76	304.62	56.98	41.12	58.17	31.75	58.49	47.38
	Ours	11.35	217.20	40.57	43.27	57.77	32.14	58.92	**48.03**
0.5	SVD	53999.48	39207.00	58558.00	25.80	47.36	25.55	52.67	37.85
	FWSVD	8173.21	8615.71	8024.67	25.84	48.70	25.64	52.83	38.25
	ASVD	6977.57	15539.44	4785.15	25.13	49.17	25.48	52.94	38.18
	SVD-LLM	272.19	1772.91	129.66	31.65	51.14	28.38	54.57	41.44
	AdaSVD	25.58	593.14	113.84	34.18	54.06	28.88	55.50	43.16
	Ours	14.02	95.48	48.94	35.56	55.80	29.58	56.58	**44.38**
0.6	SVD	65186.67	79164.00	70381.00	24.49	51.85	25.40	53.16	38.73
	FWSVD	27213.30	24962.80	47284.87	25.38	48.46	25.61	51.96	37.85
	ASVD	10003.57	15530.19	9983.83	26.68	48.86	25.76	51.80	38.28
	SVD-LLM	89.90	2052.89	561.00	26.73	47.43	26.89	53.48	38.63
	AdaSVD	50.33	1216.95	239.18	28.20	51.22	27.36	52.83	39.90
	Ours	23.89	334.67	100.42	31.02	52.01	30.38	54.62	**42.01**

Table 6: PPL and accuracy of LLaMA-3-8B across compression ratios comparing SVD-LLM and Ours.

Ratio	Method	WikiText-2 (PPL)↓	Openb↑	ARC_c↑	ARC_e↑	WinoG.↑	HellaS.↑	PIQA↑	MathQA↑	Mean↑
0.2	SVD-LLM	13.87	0.242	0.278	0.579	0.648	0.393	0.664	0.259	0.438
	Ours	**11.49**	**0.252**	**0.284**	**0.593**	**0.658**	**0.393**	**0.671**	**0.268**	**0.446**
0.4	SVD-LLM	80.46	0.134	0.193	0.325	0.523	0.274	0.548	0.208	0.315
	Ours	**23.30**	**0.162**	**0.196**	**0.340**	**0.554**	**0.296**	**0.552**	**0.218**	**0.331**
0.6	SVD-LLM	729.46	0.104	0.207	0.272	0.521	0.264	0.529	0.211	0.301
	Ours	**63.09**	**0.136**	**0.219**	**0.285**	**0.534**	**0.278**	**0.535**	**0.225**	**0.316**
0.8	SVD-LLM	6971.65	0.120	**0.207**	0.259	0.503	0.258	0.531	0.201	0.297
	Ours	**181.84**	**0.140**	0.203	**0.281**	**0.520**	**0.259**	**0.539**	**0.208**	**0.307**

We have supplemented detailed experimental results on different compression ratios for LLama-2 and LLama-3. As shown in Table 5 and Table 6, our method consistently achieves superior results across different compression ratios. Especially under extreme compression ratio conditions, such as Llama3-8B at a compression ratio of 0.8, our method outperforms the SVD-LLM method by nearly two orders of magnitude on the Perplexity metric.

A.5 Visualization of Self-Adaptation Beta

To clarify the impact of different beta coefficients on the low-rank property of the optimization objective in Section 4.1, we visualized the compression of the k-proj Linear layer of the 14th layer of LLama3-8B. As shown in Figure 5, beta affects the low-rank property of the optimization objective. The difference between the most suitable beta and the worst beta within the range of [0,1] is significant. When beta = 0.95, as shown in the first subfigure, the proportion of discarded spectral energy is 7%, while when beta = 0.4, only 4.7% of the spectral energy is lost.

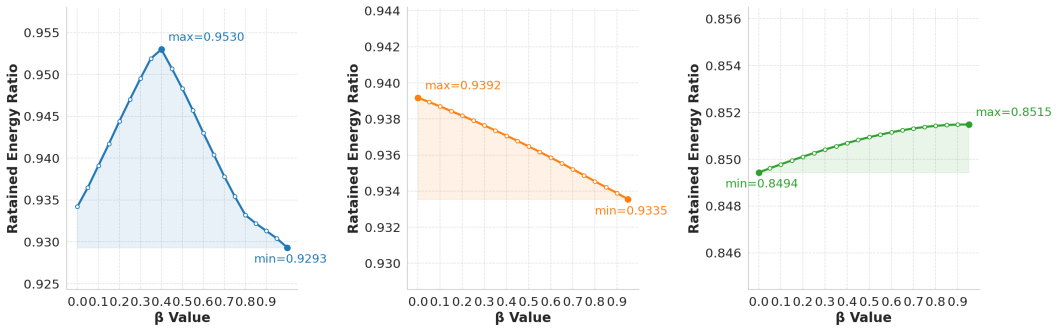

Figure 5: Using different beta coefficients for the k-proj of LLama3-8B at layer 14, layer 3 and layer 20 after the SVD in Equation 11, the proportion of retained energy at rank-1228.

We visualized in Figure 6 the Self-Adaptation beta coefficients of the odd k-proj layers of the LLama3-8B model under the condition of a compression ratio of 0.4. As can be observed from the figure, in the shallow layers of the network, beta tends to be small. As the number of layers deepens, the optimal beta continuously increases. In the last few layers, a downward trend appears.

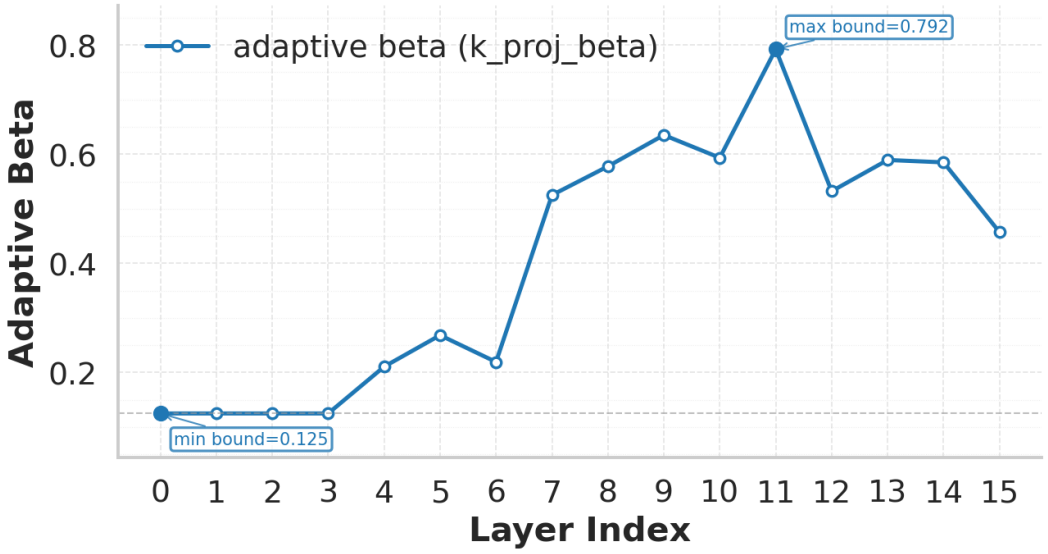

Figure 6: The visualization of the optimal beta values obtained from k-proj of odd-numbered layers on LLaMA2-7B.

A.6 ALGORITHM

APPENDIX A. ALGORITHMIC DETAILS

Algorithm 1 (CollectSecondOrderStats-Streaming). This routine gathers the per-layer second-order statistics required by SAES-SVD without storing raw activations. For each layer ℓ, it maintains a running estimate of the covariance $H_\ell = XX^\top$ and the cross-residual term $\Delta_\ell = (X^f - X)X^\top$. The update is fully streaming: existing statistics are reweighted by a factor $\gamma = n_\ell/(n_\ell + m)$ and the current mini-batch is scaled by $\sqrt{2/(n_\ell + m)}$ before accumulation, matching the numerics used in our implementation. Shapes are normalized to (d_{in}, N) to keep the math consistent across layers. This design avoids caching X_ℓ or X_ℓ^f in memory, reduces I/O, and remains robust under variable batch sizes.

Algorithm 2 (SAES-SVD for a single layer). Given layer weights W_ℓ and the statistics (H_ℓ, Δ_ℓ), we form the whitener $L_\ell = (H_\ell + \lambda I)^{-1/2}$ and the whitened objective matrix $G_\ell(\beta) = W_\ell\big(H_\ell + \beta\,\Delta_\ell\big)L_\ell$. If no alignment strength is provided, β is selected by Algorithm 3; otherwise we use the given α via $\beta = \alpha/(1+\alpha)$. A single rank-$r$ truncated SVD of $G_\ell(\beta)$ yields factors $A_\ell = \tilde{U}_\ell \Sigma_\ell^{1/2}$ and $B_\ell = \Sigma_\ell^{1/2}\tilde{V}_\ell^\top L_\ell$, which reconstruct a low-rank approximation of W_ℓ tailored to the cumulative-error–aware objective. The routine uses only one SVD per layer and a Cholesky-based whitener, making it both stable and efficient.

Algorithm 3 (ACES_BetaSelect). This procedure selects the adaptive alignment coefficient β without repeated SVDs. We compute a *single* SVD of $S = WHL$ to obtain the rank-r principal subspace, project (S, D) onto the orthogonal complement to get $(S_\perp, D_\perp)$, and then optimize a first-order (fixed-subspace) surrogate. Two objectives are supported: (i) the *ratio* objective, which minimizes the approximate tail/total energy $\widetilde{\rho}(\beta) = \frac{a+2b\beta+c\beta^2}{A+2B\beta+C\beta^2}$; and (ii) the *energy* objective, which minimizes the tail energy $a + 2b\beta + c\beta^2$. Candidate β values are obtained in closed form (stationary roots and interval endpoints), then filtered through guardrails: interval clipping $[\beta_{\min}, \beta_{\max}]$, a cap $\beta_{\text{cap}} < 1$ to prevent over-alignment, and an optional shrink factor $\rho \in (0, 1]$ for added stability. The final choice $\beta^\star$ is mapped back to $\alpha^\star = \beta^\star/(1 - \beta^\star)$ and passed to Algorithm 2.

Complexity and implementation notes. All three algorithms rely only on matrix multiplications, one Cholesky per layer for whitening, and one rank-r truncated SVD per layer (randomized or Lanczos). No iterative backpropagation or fine-tuning is required. In practice, choosing a small ridge λ ensures numerical stability when H_ℓ is ill-conditioned; if Cholesky fails, increase λ. The `ratio` objective in Algorithm 3 is recommended when cross-layer robustness is prioritized; the `energy` objective is a conservative alternative that further suppresses absolute tail energy.

Algorithm 1: CollectSecondOrderStats-Streaming (per-layer)

Input: Calibration dataset $\mathcal{D}$; layer set $\{\ell\}$; routines that expose per-layer inputs X_ℓ and FP references X_ℓ^f

Output: Per-layer second-order stats $H_\ell \in \mathbb{R}^{d_{\text{in}} \times d_{\text{in}}}$, $\Delta_\ell \in \mathbb{R}^{d_{\text{in}} \times d_{\text{in}}}$

foreach *layer* ℓ **do**
  $H_\ell \leftarrow 0, \Delta_\ell \leftarrow 0, n_\ell \leftarrow 0$ `// running matrices and sample counter`
foreach *mini-batch* $\mathcal{B} \subset \mathcal{D}$ **do**
  `// Forward once to cache both compressed-path inputs` X_ℓ `and FP references` X_ℓ^f
  run forward; collect $\{X_\ell, X_\ell^f\}_\ell$
  foreach *layer* ℓ **do**
    `// Flatten to 2D and transpose to` (d_{in}, N)
    if $\text{ndim}(X_\ell) = 2$ **then**
      $X \leftarrow X_\ell^\top$
    else if $\text{ndim}(X_\ell) = 3$ **then**
      $X \leftarrow \text{reshape}(X_\ell, -1, d_{\text{in}})^\top$
    else
      $X \leftarrow \text{flatten\_to\_2D}(X_\ell)^\top$
    Do the same for X_ℓ^f to get X^f with shape (d_{in}, N)
    `// Streaming reweighting (match code: scale old stats, add scaled current batch)`
    $t \leftarrow n_\ell, \quad m \leftarrow N, \quad \gamma \leftarrow \frac{t}{t+m}$
    $H_\ell \leftarrow \gamma H_\ell, \quad \Delta_\ell \leftarrow \gamma \Delta_\ell, \quad n_\ell \leftarrow t + m$
    `// Batch scaling by` $\sqrt{2/n_\ell}$ `(as in code)`
    $s \leftarrow \sqrt{\frac{2}{n_\ell}}$
    $\widetilde{X} \leftarrow s \cdot X, \quad \widetilde{X}^f \leftarrow s \cdot X^f$
    `// Update` $H_\ell = XX^\top$ `and` $\Delta_\ell = (X^f - X)X^\top$ `in the same metric`
    $H_\ell \leftarrow H_\ell + \widetilde{X}\,\widetilde{X}^\top$
    $dX \leftarrow \widetilde{X}^f - \widetilde{X}$
    $\Delta_\ell \leftarrow \Delta_\ell + dX\,\widetilde{X}^\top$
return $\{(H_\ell, \Delta_\ell)\}_\ell$

Algorithm 2: SAES-SVD for layer ℓ

Input: Weights $W_\ell \in \mathbb{R}^{d_{out} \times d_{in}}$; second-order statistics H_ℓ, Δ_ℓ; target rank r; ridge $\lambda \geq 0$; (optional) α or β

Output: $A_\ell \in \mathbb{R}^{d_{out} \times r}$, $B_\ell \in \mathbb{R}^{r \times d_{in}}$ s.t. $W_\ell \approx A_\ell B_\ell$

$L_\ell \leftarrow (H_\ell + \lambda \mathbf{I})^{-1/2}$ `// Whitening Matrix`
if α *is given and* β *is not* **then**
  $\beta \leftarrow \alpha/(1+\alpha)$
if β *is not given* **then**
  $\beta \leftarrow \texttt{ACES\_BetaSelect}(W_\ell, H_\ell, \Delta_\ell, r, \lambda)$
  `// Adaptive` β
$G_\ell \leftarrow W_\ell\,(H_\ell + \beta\,\Delta_\ell)\,L_\ell$ `// Whitened objective matrix`
$[\tilde{U}_\ell, \Sigma_\ell, \tilde{V}_\ell] \leftarrow \texttt{TruncatedSVD}(G_\ell, r)$ `// Trucated SVD Decomposition`
$A_\ell \leftarrow \tilde{U}_\ell\,\Sigma_\ell^{1/2}$
$B_\ell \leftarrow \Sigma_\ell^{1/2}\,\tilde{V}_\ell^\top\,L_\ell$
return (A_ℓ, B_ℓ)

Algorithm 3: ACES_BetaSelect (single-SVD, closed-form)

Input: $W \in \mathbb{R}^{m\times n}$,
$H = XX^\top$,
$\Delta = (X^f - X)X^\top$,
target rank r, damping $\lambda > 0$, interval $[\beta_{\min}, \beta_{\max}]$, cap $\beta_{\text{cap}} < 1$, shrink $\rho \in (0, 1]$, objective $\in \{\texttt{ratio}, \texttt{energy}\}$
Output: $\beta^\star \in [\beta_{\min}, \beta_{\max}]$ and $\alpha^\star = \beta^\star/(1-\beta^\star)$

Whitening and decomposition:
$L \leftarrow (H + \lambda I)^{-1/2}$ (Cholesky-based)
$S \leftarrow W\,H\,L, \quad D \leftarrow W\,\Delta\,L$ // $G(\beta) = S + \beta D$
$[U_r, \Sigma_r, V_r] \leftarrow \texttt{TopRSVD}(S, r)$ // only *one* SVD per layer
Projectors and FOA terms:
$P_L \leftarrow I - U_r U_r^\top, \quad P_R \leftarrow I - V_r V_r^\top$
$S_\perp \leftarrow P_L S P_R, \quad D_\perp \leftarrow P_L D P_R$
$a \leftarrow \|S_\perp\|_F^2,\ b \leftarrow \langle S_\perp, D_\perp\rangle,\ c \leftarrow \|D_\perp\|_F^2$
$A \leftarrow \|S\|_F^2,\ B \leftarrow \langle S, D\rangle,\ C \leftarrow \|D\|_F^2$
Candidate generation (FOA):
if *objective* = *ratio* **then**
  $p(\beta) \leftarrow (cB - bC)\beta^2 + (cA - aC)\beta + (bA - aB)$ // stationary points of $\widetilde{\rho}(\beta)$
  $\mathcal{C} \leftarrow \{\text{real roots of } p(\beta)\} \cup \{\beta_{\min}, \beta_{\max}\}$
$\beta_0 \leftarrow \text{clip}(-b/c,\ \beta_{\min},\ \beta_{\max})$ // minimizes tail energy $a + 2b\beta + c\beta^2$
$\mathcal{C} \leftarrow \{\beta_0\}$
Selection with guardrails:
foreach $\beta \in \mathcal{C}$ **do**
  $\beta \leftarrow \min\{\max\{\beta, \beta_{\min}\}, \beta_{\max}\}$
  $\beta \leftarrow \rho \cdot \min\{\beta, \beta_{\text{cap}}\}$ // shrink & cap for stability
  if *objective* = *ratio* **then**
    $\text{score}(\beta) \leftarrow \dfrac{a + 2b\beta + c\beta^2}{A + 2B\beta + C\beta^2}$ // approx. tail/total ratio $\widetilde{\rho}(\beta)$
  else
    $\text{score}(\beta) \leftarrow a + 2b\beta + c\beta^2$ // tail energy
if *objective* = *ratio* **then**
  $\beta^\star \leftarrow \arg\min_{\beta\in\mathcal{C}} \text{score}(\beta)$
else
  $\beta^\star \leftarrow \arg\min_{\beta\in\mathcal{C}} \text{score}(\beta)$
Return: $\beta^\star$ and $\alpha^\star = \beta^\star/(1-\beta^\star)$