

# ERC-SVD: Error-Controlled SVD for Large Language Model Compression

Haolei Bai<sup>1,2</sup>, Siyong Jian<sup>1,4</sup>, Tuo Liang<sup>3</sup>, Yu Yin<sup>3</sup>, Huan Wang<sup>1\*</sup>  
<sup>1</sup>Westlake University, <sup>2</sup>Nanyang Technological University,  
<sup>3</sup>Case Western Reserve University, <sup>4</sup>Nanjing University

Large language models (LLMs) have demonstrated impressive capabilities in a wide range of downstream natural language processing tasks. Nevertheless, their considerable sizes and memory demands hinder practical deployment, underscoring the importance of developing efficient compression strategies. Singular value decomposition (SVD) decomposes a matrix into orthogonal components, enabling efficient low-rank approximation. This is particularly suitable for LLM compression, where weight matrices often exhibit significant redundancy. However, current SVD-based methods neglect the residual matrix from truncation, resulting in significant truncation loss. Additionally, compressing all layers of the model results in severe error propagation. To overcome these limitations, we propose **ERC-SVD**, a new post-training SVD-based LLM compression method from an error-controlled perspective. Specifically, we leverage the residual matrix generated during the truncation process to reduce truncation loss. Moreover, under a fixed overall compression ratio, we selectively compress the last few layers of the model, which mitigates error propagation and improves compressed model performance. Comprehensive evaluations on diverse LLM families and multiple benchmark datasets indicate that ERC-SVD consistently achieves superior performance over existing counterpart methods, demonstrating its practical effectiveness.

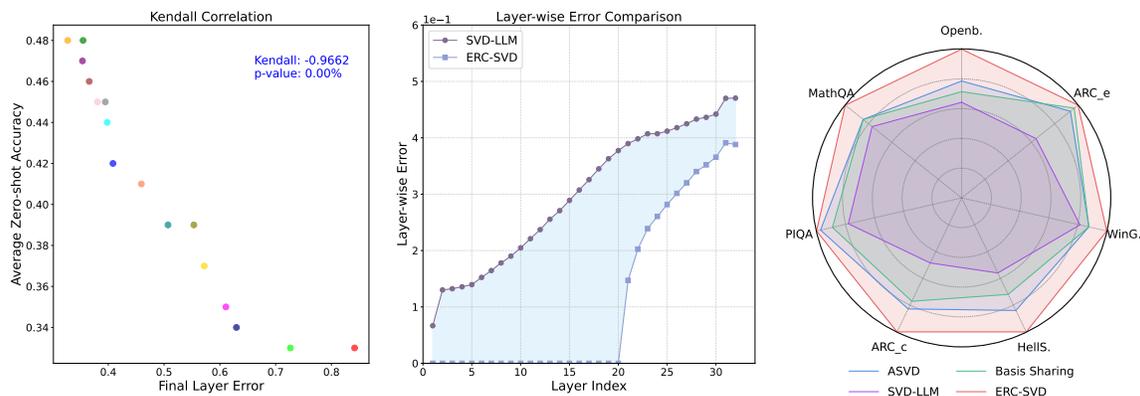


Figure 1: Left: The Kendall correlation between the final-layer error and the average zero-shot accuracy. Middle: Layer-wise error comparison between ERC-SVD and SVD-LLM of LLaMA-2-7B under 20% compression ratio. Right: The accuracy comparison of LLaMA-2-7B compressed by different methods under 20% compression ratio on seven reasoning and understanding tasks.

## 1. Introduction

Large language models (LLMs) have emerged as powerful tools, delivering state-of-the-art performance across a wide range of tasks such as text generation, translation, and reasoning. The scaling law [1] has driven a trend toward increasingly large models, exemplified by models such as GPT [2], PaLM [3], LLaMA [4], Deepseek [5], and Qwen [6], which often contain tens to hundreds

\*Corresponding author: wanghuan@westlake.edu.cn

of billions of parameters. Despite their powerful capabilities, the enormous scale of LLMs poses serious challenges for efficient deployment due to high computational demands [7–9]. This resource burden not only limits deployment on edge devices and consumer-level hardware but also increases the cost and carbon footprint of serving LLMs in production [10, 11].

As the scale of LLMs continues to grow, compression techniques, including weight quantization [12–15], network pruning [16–19], knowledge distillation [20–23], and low-rank decomposition [24–28], have become increasingly important for the practical deployment of LLMs in resource-constrained environments. The syntactic and semantic correlations acquired during training often induce redundancy in LLM weight matrices, giving rise to a low-rank structure [29]. As a result, singular value decomposition (SVD) provides a principled and effective approach for compressing these matrices with minimal loss. In particular, post-training approaches are gaining traction, as they can significantly reduce memory and compute requirements without the need for expensive retraining, making them especially suitable for scaling up foundation models. Recent studies on post-training SVD-based LLM compression, including ASVD [26], SVD-LLM [27], Basis Sharing [30], and AdaSVD [28], have made significant progress in reducing model size while preserving performance, demonstrating the effectiveness of low-rank approximation techniques in compressing LLMs. However, existing methods suffer from two major limitations. First, existing methods ignore the importance of the residual matrix generated during SVD truncation, leading to significant truncation loss. Second, compressing all model layers often results in high layer-wise error and severe error propagation.

In this work, we propose **ERC-SVD**, a new post-training SVD-based compression method for LLMs. Building upon the key observations outlined above, ERC-SVD introduces two core technical innovations. ① **Residual compensation for SVD truncation**: The residual matrix produced during SVD truncation can be effectively utilised to reduce the truncation loss. Specifically, we perform SVD truncation in two stages: we first truncate the original weight matrix  $\mathbf{W}$  to obtain its intermediate low-rank approximation  $\mathbf{W}_{r_i}$ . After that, we compute the residual matrix  $\mathbf{R}$  between  $\mathbf{W}$  and  $\mathbf{W}_{r_i}$ . Second, we apply SVD truncation to  $\mathbf{R}$ , yielding  $\mathbf{R}_{r_r}$ . Finally, we construct the compressed weight matrix  $\hat{\mathbf{W}}_r = \mathbf{W}_{r_i} + \mathbf{R}_{r_r}$ , where the rank satisfies  $r_i + r_r = r$ . Detailed description and a mathematical proof are provided in Section 3.1. ② **Partial-layer compression for SVD**: LLMs consist of a sequence of consecutive layers, where the output of each layer serves as the input to the next. Therefore, any error introduced in earlier layers can propagate and accumulate through subsequent layers, leading to severe performance degradation. To mitigate this, we propose compressing only the last few layers under a fixed overall compression ratio while keeping the earlier layers intact. This strategy ensures that the earlier layers remain error-free, reducing the impact of error propagation.

Our key contributions can be summarised as follows:

- We introduce **residual compensation for SVD truncation**, a theoretically grounded compensation strategy. By leveraging the residual matrix generated during the SVD truncation, our strategy significantly reduces the overall truncation loss.
- We propose **partial-layer compression for SVD**, which compresses only the last few layers of the model under a fixed overall compression ratio. This strategy effectively reduces layer-wise error and mitigates error propagation.
- Extensive experiments on multiple LLMs (LLaMA, OPT, Mistral, Vicuna, and Qwen) and benchmark datasets (both language modelling and zero-shot reasoning) demonstrate that ERC-SVD outperforms existing methods across a wide range of compression ratios.

## 2. Related Work

### 2.1. Techniques for Large Language Model Compression

The growing size of large language models (LLMs) has raised increasing concerns over their computational and memory demands. To address these challenges, a variety of model compression techniques have been proposed. Conventional approaches often require computationally expensive

retraining, which is generally impractical due to the substantial computational cost associated with the massive size of LLMs. Consequently, recent efforts have shifted toward more resource-friendly post-training compression techniques [12, 17, 31]. Typically employed approaches include network pruning and weight quantization. And pruning techniques can be classified into unstructured and structured methods. Unstructured pruning removes individual weights based on importance scores. For example, SparseGPT [17] performs one-shot pruning using second-order approximations without retraining. However, since unstructured pruning retains the original matrix shape, it offers limited inference acceleration and requires specialized hardware. In contrast, structured pruning eliminates entire blocks or channels, enabling compatibility with conventional hardware platforms. LLM-Pruner [16] groups dependent linear projections into coupled structures, assigns each group a loss-aware importance score, prunes the least important groups, and applies LoRA fine-tuning to restore performance. Additionally, ZipLM [32] prioritizes pruning components that yield the worst trade-off between latency and loss, but this often causes notable performance degradation. Quantization provides another mainstream solution. GPTQ [12] gradually quantizes and updates each weight using the Hessian matrix to minimize the quantization error. AWQ [13] preserves important weight channels by selecting reparameterization coefficients via grid search. Furthermore, BiLLM [14] and ARB-LLM [15] push quantization to the 1-bit level, while still delivering impressive downstream task accuracy. However, these techniques still cause significant performance degradation, especially at low bit-widths, due to the lack of weight and activation adaptation.

## 2.2. SVD-based Techniques for Compressing LLMs

Singular value decomposition (SVD) is a commonly employed technique for reducing matrix dimensionality by representing the original matrix as the product of two low-rank factor matrices. Recent research has demonstrated the potential of SVD-based LLM compression methods, yet comprehensive exploration remains limited. Hsu et al. [24] propose FWSVD, incorporating Fisher information to weight the importance of parameters. However, it relies on computationally intensive training and was originally applied only to small language models (e.g., BERT [33], ALBERT [34]). LoRD [25] first applies SVD to LLMs by grouping layers to improve efficiency. However, LoRD overlooks the importance of input activations. To address this limitation, Yuan et al. [26] introduce ASVD, which mitigates the impact of activation outliers by reshaping the weight matrix based on activation distribution, thereby enhancing the precision of the low-rank decomposition. CALDERA [29] exploits the inherent low-rank structure of weight matrices by approximating them through a low-rank, low-precision decomposition. Additionally, SVD-LLM [27] establishes a direct mapping between singular values and truncation loss, which means that truncating the smallest singular values leads to minimal truncation loss. Basis Sharing [30] achieves more effective LLM compression through sharing parameters across different layers. More recently, AdaSVD [28] proposes an adaptive compensation method, iteratively updating matrices during truncation.

Despite these advances, all existing methods ignore the residual matrix produced during SVD truncation, which can significantly compensate for the SVD truncation loss. Moreover, compressing only the last few layers of the model under a fixed target compression ratio can provide better performance for compressed models. However, existing methods compress all layers, either apply a uniform compression ratio across all layers or assign variable ratios based on layer-wise importance, which often results in sub-optimal performance.

## 3. ERC-SVD

The framework of ERC-SVD is illustrated in Figure 2. We first perform SVD to compress the last few layers of the model, while ensuring that the overall compression ratio is satisfied, and compute the corresponding residual matrix. A second SVD is then applied to the residual matrix to obtain its low-rank approximation. The two truncated matrices are subsequently combined to construct the final compressed weight matrices. Compared with other methods [26–28, 30], we utilize the residual matrix during the truncation process to compensate for the truncation loss. In Section 3.1, we first

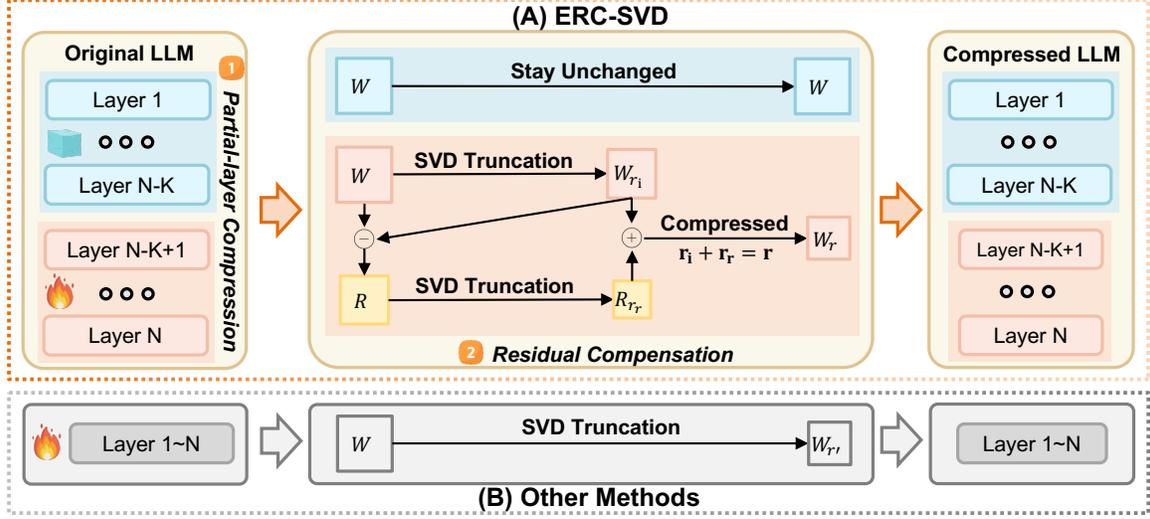


Figure 2: The overall framework of ERC-SVD, and comparison with other methods. The last  $k$  layers are selected through **partial-layer compression** and compressed using **residual compensation** with calibration data. ■ denotes these layers remain intact, while ■ denotes these layers are replaced by low-rank approximations. The overall compression ratio is  $R_o$ , for ERC-SVD, the first  $(N - k)$  layers stay unchanged, and the layer compression ratio  $R_l$  for last  $k$  layers is  $(N \cdot R_o)/k$ .

provide a detailed description of how **residual compensation** effectively works. Subsequently, in Section 3.2, we explain the benefits of **partial-layer compression** in mitigating error propagation. The pseudocode of ERC-SVD is shown in Algorithm 1, with the pseudocode for residual compensation and partial-layer compression provided in Algorithm 2 and Algorithm 3, respectively.

### 3.1. Residual Compensation for SVD Truncation

**Preliminaries for SVD.** SVD-based compression methods apply SVD to the original weight matrix  $\mathbf{W} \in \mathbb{R}^{m \times n}$ , discarding the smallest singular values to obtain its low-rank approximation  $\mathbf{W}_r$ :

$$\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \approx \mathbf{U}_r\mathbf{\Sigma}_r\mathbf{V}_r^T = \mathbf{W}_r, r < \min(m, n), \quad (1)$$

where  $\mathbf{U}_r \in \mathbb{R}^{m \times r}$  and  $\mathbf{V}_r^T \in \mathbb{R}^{r \times n}$  are composed of the top- $r$  left and right singular vectors, respectively, and  $\mathbf{\Sigma}_r \in \mathbb{R}^{r \times r}$  is a diagonal matrix containing the corresponding singular values. Following prior post-training SVD-based works [26, 27], the optimization objective for SVD truncation in LLMs can be formulated as:

$$\hat{\mathbf{W}}_r = \arg \min_{\mathbf{W}_r} \|\mathbf{W}\mathbf{X} - \mathbf{W}_r\mathbf{X}\|_F, \quad (2)$$

$$\mathcal{L} = \|\mathbf{W}\mathbf{X} - \mathbf{W}_r\mathbf{X}\|_F = \|(\mathbf{W} - \mathbf{W}_r)\mathbf{X}\|_F, \quad (3)$$

where  $\mathbf{X}$  denotes the activation of  $\mathbf{W}$  given an input, and  $\mathcal{L}$  is the truncation loss measured by the Frobenius norm. Although previous works [24, 26, 27] have made significant progress in minimizing  $\mathcal{L}$ , they consistently overlook the residual matrix generated. Based on Equation 3, minimizing the truncation loss reduces to minimizing the discrepancy between the original weight matrix  $\mathbf{W}$  and its low-rank approximation  $\mathbf{W}_r$ . Accordingly, we reformulate the optimization objective as:

$$\hat{\mathbf{W}}_r = \arg \min_{\mathbf{W}_r} \|\mathbf{W} - \mathbf{W}_r\|. \quad (4)$$

**Explanation for Residual Compensation.** Given the original weight matrix  $\mathbf{W} \in \mathbb{R}^{m \times n}$  and the layer compression ratio  $R_l$ , the compression ratio in this work denotes the fraction of parameters removed. We define the scale of the matrix as:  $\alpha = (m \cdot n)/(m + n)$ , where  $m$  and  $n$  represent the

---

**Algorithm 1** Pseudocode of ERC-SVD
 

---

**Input:** Original LLM:  $M$ , weight matrix:  $\mathbf{W} \in \mathbb{R}^{m \times n}$ , number of model layers:  $N$ , residual compensation rank  $r_r$ , step:  $s$

**Output:** Compressed LLM  $M'$  by ERC-SVD

- 1:  $CD \leftarrow$  Randomly collect calibration samples from the dataset
- 2:  $\text{Set}_{\mathbf{W}} \leftarrow M, \text{Set}_{\mathbf{W}'} \leftarrow \emptyset$  ▷ Initialize the sets of weight matrices
- 3:  $k, R_l \leftarrow \text{PARTIAL-LAYER COMPRESSION}(M, N, R_o, s)$
- 4: **for** Layer  $i$  in original LLM  $M$  **do**
- 5:   **if**  $i \in [1, N - k]$  **then**
- 6:      $\text{Set}_{\mathbf{W}'}(i) \leftarrow \text{Set}_{\mathbf{W}}(i)$  ▷ Current weight matrices stay the same
- 7:   **else**
- 8:      $\mathbf{W}_i \leftarrow \text{Set}_{\mathbf{W}}(i)$  ▷ Initialize weight matrices to compress
- 9:      $\text{Set}_{\mathbf{W}'}(i) \leftarrow \text{RESIDUAL COMPENSATION}(M, \mathbf{W}_i, R_l, r_r)$
- 10:   **end if**
- 11:    $\text{Set}_{\mathbf{W}'} \leftarrow \text{Set}_{\mathbf{W}'}(i) \cup \text{Set}_{\mathbf{W}'}$  ▷ Append weight matrices after operation
- 12: **end for**
- 13:  $M' \leftarrow \text{UPDATE}(M, CD, \text{Set}_{\mathbf{W}'})$
- 14: **return**  $M'$

---

input and output dimensions of the matrix, hence, the target rank for each layer is  $r = (1 - R_l) \cdot \alpha$ . We introduce a residual compensation factor  $\beta$ , a hyperparameter fixed at 0.05 in all experiments. The target rank  $r$  is decomposed into two components: a residual compensation rank  $r_r$ , defined as  $r_r = \alpha \cdot \beta$ , and an intermediate rank  $r_i$ , defined as  $r_i = r - r_r$ .

The entire compression process comprises two SVD truncations. First, we apply SVD to  $\mathbf{W}\mathbf{S}$ , where  $\mathbf{S}$  is the scaling weight matrix obtained from the input activation, and retain the top- $r_i$  singular values:  $\mathbf{W}\mathbf{S} \approx \mathbf{W}'_{r_i} = \mathbf{U}_{r_i} \Sigma_{r_i} \mathbf{V}'_{r_i}{}^T$ , and multiply  $\mathbf{S}^{-1}$  to obtain an intermediate low-rank approximation of the original weight matrix:  $\mathbf{W}_{r_i} = \mathbf{W}'_{r_i} \mathbf{S}^{-1} = \mathbf{U}_{r_i} \Sigma_{r_i} \mathbf{V}_{r_i}{}^T$ . The residual matrix  $\mathbf{R}$  is computed as the difference between  $\mathbf{W}$  and  $\mathbf{W}_{r_i}$ , where  $\mathbf{R} = \mathbf{W} - \mathbf{W}_{r_i}$ . We then perform a second SVD on  $\mathbf{R}$  and retain its top- $r_r$  singular values:  $\mathbf{R}_{r_r} = \mathbf{U}_{r_r} \Sigma_{r_r} \mathbf{V}_{r_r}{}^T$ . Finally, we construct the compressed weight matrix by combining the two approximations:

$$\hat{\mathbf{W}}_r = \mathbf{W}_{r_i} + \mathbf{R}_{r_r} = \mathbf{U}_r \Sigma_r \mathbf{V}_r{}^T = \hat{\mathbf{U}}_r \hat{\mathbf{V}}_r{}^T. \quad (5)$$

In this way, the residual-compensated compressed matrix  $\hat{\mathbf{W}}_r$  provides a closer approximation to the original weight matrix  $\mathbf{W}$  than the directly truncated matrix  $\mathbf{W}_r$ , as proven in the following.

**Lemma 1.** *Eckart-Young-Mirsky Theorem [35]. Let  $\mathbf{A} \in \mathbb{R}^{m \times n}$  be a matrix and  $r \leq \min(m, n)$  be a given rank. If  $\mathbf{A}_r$  denotes the optimal rank- $r$  approximation of  $\mathbf{A}$  by SVD, then for any matrix  $\mathbf{B}$  of rank  $r$ , the following inequality holds, demonstrating that  $\mathbf{A}_r$  is the optimal rank- $r$  approximation of  $\mathbf{A}$ :*

$$\|\mathbf{A} - \mathbf{A}_r\|_F^2 \leq \|\mathbf{A} - \mathbf{B}\|_F^2. \quad (6)$$

**Lemma 2.** *Let  $\mathbf{W}'$  be the rank- $r$  approximation defined by the weighted low-rank reconstruction  $\mathbf{W}' = \text{SVD}_r(\mathbf{W}\mathbf{S})\mathbf{S}^{-1}$ . A direct consequence of Lemma 1 is the following inequality:*

$$\|\mathbf{W} - \text{SVD}_r(\mathbf{W})\|_F^2 \leq \|\mathbf{W} - \mathbf{W}'\|_F^2, \quad (7)$$

**Theorem 3.** *If  $\mathbf{W}_r$  is the rank- $r$  direct truncation of the original matrix  $\mathbf{W}$ , and  $\hat{\mathbf{W}}_r$  is the residual-compensated compressed matrix, where  $r = r_i + r_r$ . Then, under the weight reconstruction error, ERC-SVD is superior to the direct truncation method.*

*Proof.* The direct truncation method extracts the first  $r$  components from the  $\mathbf{W}\mathbf{S}$  at once. We can decompose it into two parts: the **first  $r_i$  components** and the **last  $r_r$  components**:

$$\mathbf{W}_r = \underbrace{\text{SVD}_{r_i}(\mathbf{W}\mathbf{S})\mathbf{S}^{-1}}_{\mathbf{W}_{r_i}} + \underbrace{(\text{SVD}_r(\mathbf{W}\mathbf{S}) - \text{SVD}_{r_i}(\mathbf{W}\mathbf{S}))\mathbf{S}^{-1}}_{\Delta} \quad (8)$$

For ERC-SVD, the truncation loss is given by:

$$\|\mathbf{W} - \hat{\mathbf{W}}_r\|_F^2 = \|\mathbf{W} - (\mathbf{W}_{r_i} + \mathbf{R}_{r_r})\|_F^2 = \|\mathbf{R} - \mathbf{R}_{r_r}\|_F^2. \quad (9)$$

Given that  $\mathbf{R}_{r_r}$  is obtained by direct SVD, which is the optimal rank- $r_r$  approximation of  $\mathbf{R}$ . According to Lemma 1, for any matrix  $\mathbf{X}$  with rank  $r_r$ , we have:

$$\|\mathbf{R} - \mathbf{R}_{r_r}\|_F^2 \leq \|\mathbf{R} - \mathbf{X}\|_F^2. \quad (10)$$

For the direct truncation method, the truncation loss is given by:

$$\|\mathbf{W} - \mathbf{W}_r\|_F^2 = \|\mathbf{W} - (\mathbf{W}_{r_i} + \Delta)\|_F^2 = \|\mathbf{R} - \Delta\|_F^2. \quad (11)$$

$\mathbf{R}_{r_r}$  is the optimal rank- $r_r$  approximation, whereas  $\Delta$  is an approximation derived from the weighted space  $\mathbf{W}\mathbf{S}$ . Hence, by combining Lemma 1 and Inequality 10, the following relationship holds:

$$\|\mathbf{R} - \mathbf{R}_{r_r}\|_F^2 \leq \|\mathbf{R} - \Delta\|_F^2. \quad (12)$$

By substituting these results into the original form, we obtain the final inequality:

$$\boxed{\|\mathbf{W} - \hat{\mathbf{W}}_r\|_F^2 \leq \|\mathbf{W} - \mathbf{W}_r\|_F^2} \quad (13)$$

Therefore, the residual compensation strategy provides a closer approximation to the original weight matrix  $\mathbf{W}$  than direct truncation methods.  $\square$

### 3.2. Partial-layer Compression for SVD

Previous works [26–28, 30] compress all model layers even if they assign layer-specific ratios based on their relative importance, which often leads to a high layer-wise error, resulting in noticeable performance degradation. We compare the layer-wise error of LLMs across four families with different layer selection strategies; the results of LLaMA-7B and OPT-6.7B are shown in Figure 3, the results of other LLM models are presented in Figure 4. There is a significant error propagation across multiple LLM families, where the error progressively accumulates layer by layer during the forward pass. This phenomenon is particularly pronounced when compressing only the first few layers. A higher layer compression ratio must be applied to these layers, with the constraint of the overall compression ratio. Compressing the model at such an early stage introduces substantial approximation errors, which are then propagated through the remaining layers. Although the later layers are left uncompressed, the forward pass carries the accumulated error through the network. Consequently, in all evaluated LLM families, these early compressed layers exhibit the highest layer-wise error across model layers. In contrast, when only compressing the last few layers, the earlier layers introduce zero error. Although the relatively high compression ratio leads to slightly faster error accumulation, the overall error remains significantly lower than that of compressing all layers.

Interestingly, although the final-layer error converges to a narrow range regardless of how many the last layers are compressed, differences within this range still affect model performance. As illustrated in the left figure of Figure 1, there is a strong correlation between the final-layer error and the average zero-shot accuracy. Motivated by this observation, we select the number of last layers to compress by minimizing the final-layer error, and the layer compression ratio  $R_l$  must satisfy the following mathematical relation to ensure a fair comparison:

$$R_l = (N \cdot R_o)/k, k \in \{1, 2, \dots, N - 1 \mid R_l < 1\}, \quad (14)$$

where  $N$  is the number of model layers,  $R_o$  represents the overall compression ratio. As shown in the middle figure of Figure 1, compared to SVD-LLM [27], our strategy significantly reduces layer-wise error across all model layers and mitigates error propagation.

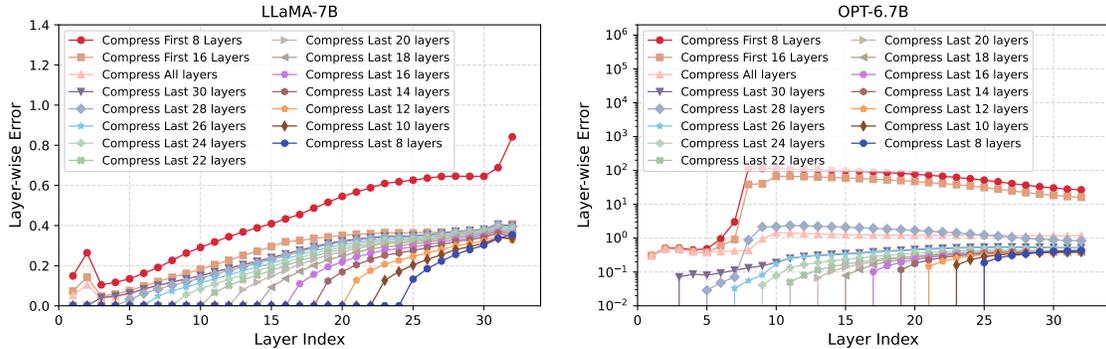


Figure 3: Layer-wise error comparison between the original model, LLaMA-7B, and OPT-6.7B compressed by ERC-SVD with different layer selection strategies on WikiText-2. The overall compression ratio is 20%, and all layer selection strategies strictly adhere to the compression constraint.

## 4. Experiments

### 4.1. Setups

**Evaluation Details.** We compare our method against five baselines without re-training: conventional SVD, ASVD [26], SVD-LLM [27], Basis Sharing [30], and AdaSVD [28]. We evaluate ERC-SVD on ten models spanning five LLM families: LLaMA-2-7B [36], LLaMA-13B [4], LLaMA-2-13B [36], LLaMA-30B [4], OPT-6.7B [37], OPT-13B [37], OPT-30B [37], Mistral-7B [38], Vicuna-7B [39], Qwen-3-8B [6]. For language modeling, we use three benchmark datasets: WikiText-2 [40], PTB [41], and C4 [42]. For zero-shot reasoning and understanding, we evaluate on seven tasks within the *LM-Evaluation-Harness* framework: OpenbookQA [43], WinoGrande [44], HellaSwag [45], PIQA [46], MathQA [47], ARC\_e and ARC\_c [48]. Moreover, to demonstrate the generality and effectiveness of ERC-SVD, we further extend our evaluation to vision-language models (VLMs), LLaVA [49].

**Implementation Details.** We follow the protocols of ASVD, SVD-LLM, and Basis Sharing, randomly selecting 256 calibration samples with a sequence length of 2048 from WikiText-2. We then apply matrix scaling prior to performing SVD truncation. All methods are implemented with PyTorch and Transformers on NVIDIA A100 GPUs.

### 4.2. Results

We evaluate the overall performance of ERC-SVD across four dimensions: ① **Effectiveness under different compression ratios** (ranging from 20% to 60%), ② **Scalability to larger-scale models**, ③ **Generalizability across diverse LLM families**, and ④ **Performance on VLMs**, including quantitative and qualitative results, are provided in Table 14 and Appendix A.8, respectively.

**Performance under different compression ratios.** We evaluate the performance of LLaMA-2-7B [36] compressed by ERC-SVD, conventional SVD, and existing post-training baselines under compression ratios ranging from 20% to 60% across ten benchmark datasets. The results for LLaMA-2-7B are presented in Table 1. Across three language modeling datasets, WikiText-2 [40], PTB [41], and C4 [42], ERC-SVD consistently outperforms all baselines across most evaluated compression ratios, with only slight suboptimality observed in specific cases. In particular, on PTB and C4, the improvements are more pronounced, suggesting that ERC-SVD exhibits stronger generalization capability. In addition, across seven common sense reasoning datasets, ERC-SVD outperforms the strongest existing baseline on most datasets, achieving the highest average accuracy across all evaluated compression ratios and further demonstrating its robustness and overall effectiveness.

**Performance on larger-scale LLMs.** To examine the scalability of ERC-SVD, we evaluate its performance on larger-scale LLMs from two representative families: LLaMA and OPT (13B and 30B).

Table 1: Overall performance of LLaMA-2-7B [36] compressed by ERC-SVD and baselines under 20% to 60% compression ratios (“RATIO”), including performance on three language modeling datasets (measured by perplexity ( $\downarrow$ )) and zero-shot performance on seven common sense reasoning datasets (measured by individual and average accuracy ( $\uparrow$ )). The best results are marked in **bold**. NaN denotes evaluation failure due to numerical instability. \* refers to results derived from the original paper. - means that results are not available.

RATIO	METHOD	WikiText-2 $\downarrow$	PTB $\downarrow$	C4 $\downarrow$	Openb. $\uparrow$	ARC_e $\uparrow$	WinoG. $\uparrow$	HellaS. $\uparrow$	ARC_c $\uparrow$	PIQA $\uparrow$	MathQA $\uparrow$	Average $\uparrow$
-	Original	5.47	26.84	7.28	0.31	0.69	0.67	0.56	0.40	0.78	0.27	0.53
20%	SVD	18208.79	59320.78	27131.56	0.14	0.26	0.49	0.25	0.22	0.52	0.20	0.30
	ASVD	9.56	120.74	<b>12.85</b>	<b>0.25</b>	0.59	<b>0.61</b>	<b>0.46</b>	<b>0.32</b>	<b>0.72</b>	<b>0.24</b>	<b>0.45</b>
	SVD-LLM	8.37	139.68	20.18	0.23	0.50	0.59	0.39	0.26	0.65	0.23	0.41
	Basis Sharing	<u>7.77</u>	<u>60.00</u>	16.22	0.24	<b>0.60</b>	<b>0.61</b>	0.43	0.31	0.69	<u>0.24</u>	<u>0.45</u>
	<b>ERC-SVD</b>	<b>7.63</b>	<b>45.37</b>	<b>14.73</b>	<b>0.28</b>	<b>0.61</b>	<b>0.65</b>	<b>0.50</b>	<b>0.35</b>	<b>0.73</b>	<b>0.26</b>	<b>0.48</b>
30%	SVD	30373.39	48930.94	36905.54	0.12	0.25	0.49	0.25	0.22	0.51	0.21	0.29
	ASVD	984.03	NaN	NaN	0.15	0.27	0.51	0.26	0.22	0.53	0.20	0.31
	SVD-LLM	10.66	292.90	34.96	0.21	0.42	0.55	0.34	0.22	0.60	<b>0.23</b>	0.37
	Basis Sharing	<b>9.69</b>	<b>101.22</b>	<b>25.45</b>	<b>0.23</b>	<b>0.54</b>	<b>0.59</b>	<b>0.38</b>	<b>0.27</b>	<b>0.64</b>	<b>0.23</b>	<b>0.41</b>
	<b>ERC-SVD</b>	<b>10.32</b>	<b>73.04</b>	<b>23.28</b>	<b>0.23</b>	<b>0.51</b>	<b>0.62</b>	<b>0.42</b>	<b>0.29</b>	<b>0.68</b>	<b>0.24</b>	<b>0.43</b>
40%	SVD	39524.00	68829.98	56518.95	0.13	0.26	0.50	0.25	0.21	0.52	0.18	0.29
	ASVD	NaN	NaN	NaN	0.15	0.25	0.50	0.26	0.22	0.52	0.18	0.30
	SVD-LLM	16.11	717.34	61.96	0.16	0.35	0.55	0.30	0.20	0.57	<b>0.23</b>	0.34
	Basis Sharing	<b>14.66</b>	<b>207.49</b>	<b>47.41</b>	0.18	<b>0.43</b>	<b>0.57</b>	<b>0.33</b>	0.22	<b>0.61</b>	<b>0.23</b>	<b>0.37</b>
	AdaSVD*	14.76	304.62	56.98	<b>0.19</b>	0.41	<b>0.58</b>	0.32	<b>0.23</b>	0.58	-	-
<b>ERC-SVD</b>	<b>14.17</b>	<b>136.32</b>	<b>43.19</b>	<b>0.20</b>	<b>0.43</b>	<b>0.57</b>	<b>0.35</b>	<b>0.24</b>	<b>0.63</b>	<b>0.23</b>	<b>0.38</b>	
50%	SVD	53405.48	39023.05	58547.82	0.15	0.25	0.48	0.25	0.22	0.53	0.18	0.29
	ASVD	NaN	NaN	NaN	0.13	0.26	0.50	0.25	<b>0.23</b>	0.50	0.20	0.30
	SVD-LLM	27.19	1775.52	129.71	<b>0.14</b>	0.30	0.50	0.28	0.20	0.54	0.23	0.31
	Basis Sharing	<b>23.78</b>	689.03	115.84	0.15	<b>0.34</b>	<b>0.54</b>	<b>0.29</b>	0.20	<b>0.56</b>	<b>0.23</b>	<b>0.33</b>
	AdaSVD*	25.58	<b>593.14</b>	<b>113.84</b>	<b>0.15</b>	<b>0.34</b>	<b>0.54</b>	<b>0.29</b>	0.20	<b>0.56</b>	-	-
<b>ERC-SVD</b>	<b>24.26</b>	<b>286.24</b>	<b>100.34</b>	<b>0.14</b>	<b>0.35</b>	<b>0.55</b>	<b>0.31</b>	<b>0.22</b>	<b>0.59</b>	<b>0.22</b>	<b>0.34</b>	
60%	SVD	65240.23	79002.21	70659.74	<b>0.14</b>	0.25	0.50	0.25	0.23	0.52	0.19	0.30
	ASVD	NaN	19581.17	NaN	<b>0.15</b>	0.25	0.50	0.25	<b>0.23</b>	0.52	0.12	0.29
	SVD-LLM	<b>54.19</b>	3442.74	<b>263.02</b>	<b>0.14</b>	0.26	0.50	<b>0.27</b>	0.20	<b>0.53</b>	<b>0.21</b>	<b>0.30</b>
	Basis Sharing	<b>50.59</b>	<b>2055.63</b>	299.36	<b>0.14</b>	<b>0.29</b>	<b>0.51</b>	<b>0.27</b>	0.20	<b>0.52</b>	<b>0.21</b>	<b>0.30</b>
	AdaSVD*	60.08	2137.28	294.26	0.12	<b>0.27</b>	0.50	0.27	0.20	<b>0.53</b>	-	-
<b>ERC-SVD</b>	<b>68.59</b>	<b>991.48</b>	<b>255.70</b>	0.13	<b>0.29</b>	<b>0.52</b>	<b>0.28</b>	<b>0.21</b>	<b>0.55</b>	<b>0.22</b>	<b>0.31</b>	

**Performance on Multiple LLM families.** To evaluate the generalization ability of ERC-SVD, we apply it to three LLMs from different families, including OPT-6.7B [37], Mistral-7B [38], and Vicuna-7B [39]. As shown in Table 2, under 30% compression ratio, ERC-SVD consistently outperforms all baselines on three language modeling benchmarks across these architectures. Notably, the largest overall improvements are observed on Mistral-7B, with perplexity reductions of 71% on WikiText-2, 45% on PTB, and 46% on C4. We reproduce ASVD and SVD-LLM using their public codebases. While ASVD fails in certain cases due to numerical instability (denoted as NaN in the table), ERC-SVD consistently maintains stable and reliable performance. Additional zero-shot accuracy results comparison about these LLMs are provided in Table 9.

Table 2: Perplexity ( $\downarrow$ ) of different LLM structures under 30% compression ratio.

MODEL	METHOD	WikiText-2 $\downarrow$	PTB $\downarrow$	C4 $\downarrow$
OPT-6.7B	SVD	116067.28	86760.50	168165.89
	ASVD	<b>26.67</b>	71.36	44.51
	SVD-LLM	28.03	<b>37.46</b>	<b>40.35</b>
	<b>ERC-SVD</b>	<b>17.10</b> ( <b>↓36%</b> )	<b>27.24</b> ( <b>↓27%</b> )	<b>38.40</b> ( <b>↓45%</b> )
Mistral-7B	SVD	59569.54	57830.63	78168.24
	ASVD	221.66	927.15	266.04
	SVD-LLM	<b>48.94</b>	<b>193.22</b>	<b>56.55</b>
	<b>ERC-SVD</b>	<b>14.09</b> ( <b>↓71%</b> )	<b>105.37</b> ( <b>↓45%</b> )	<b>30.72</b> ( <b>↓46%</b> )
Vicuna-7B	SVD	24835.33	24510.90	29368.55
	ASVD	106.32	NaN	NaN
	SVD-LLM	<b>12.42</b>	<b>104.27</b>	<b>39.55</b>
	<b>ERC-SVD</b>	<b>11.57</b> ( <b>↓7%</b> )	<b>69.28</b> ( <b>↓34%</b> )	<b>27.24</b> ( <b>↓31%</b> )

Table 3: Perplexity ( $\downarrow$ ) of larger-scale LLMs under 20% compression ratio on PTB [41].

METHOD	LLaMA-13B	LLaMA-30B	LLaMA-2-13B	OPT-13B	OPT-30B
SVD	1878.04	555.55	5464.57	1552.55	250.49
ASVD	12.42	31.36	81.00	36.47	30.95
SVD-LLM	<b>12.17</b>	<b>9.10</b>	88.13	<b>14.86</b>	<b>12.94</b>
<b>ERC-SVD</b>	<b>9.70</b>	<b>8.41</b>	<b>66.47</b>	<b>13.22</b>	<b>12.89</b>

As presented in Table 3, ERC-SVD consistently achieves superior performance over existing baselines under 20% compression ratio, demonstrating robust effectiveness across varying model scales. Additional zero-shot accuracy results about these larger-scale LLMs are provided in Table 6.

### 4.3. Ablation Study

We present several ablation studies to assess the robustness of ERC-SVD. ① **Effectiveness of residual compensation (REC) and partial-layer compression (PLC):** We assess the individual contributions of REC and PLC. ② **Impact of residual compensation factor:** We conduct experiments to examine

Table 4: Ablation results of REC and PLC on LLaMA-2-7B under 30% and 40% compression ratios.

RATIO	METHOD	REC	PLC	C4↓	Openb.↑	ARC_e↑	WinoG.↑	HellaS.↑	ARC_c↑	PIQA↑	MathQA↑	Average↑
30%	ASVD	-	-	NaN	0.15	0.27	0.51	0.26	0.22	0.53	0.20	0.31
	SVD-LLM	-	-	34.96	0.21	0.42	0.55	0.34	0.22	0.60	0.23	0.37
	ERC-SVD	✓	✗	30.68	0.22	0.42	0.58	0.35	0.23	0.61	0.24	0.38
		✗	✓	24.77	0.22	0.48	0.60	0.38	0.25	0.65	0.22	0.40
		✓	✓	<b>23.28</b>	0.23	0.51	0.62	0.42	0.29	0.68	0.24	<b>0.43</b>

Table 5: Perplexity (↓) and zero-shot individual and average accuracy (↑) of LLaMA-2-7B 30% compressed by ERC-SVD with different residual compensation factors ( $\beta$ ).

$\beta$	WikiText-2↓	Openb.↑	ARC_e↑	WinoG.↑	HellaS.↑	ARC_c↑	PIQA↑	MathQA↑	Average↑
0.025	10.38	0.23	0.51	0.62	0.41	0.28	0.67	0.23	0.42
0.050	10.32	0.23	0.51	0.62	0.42	0.29	0.68	0.24	<b>0.43</b>
0.075	<b>10.11</b>	0.22	0.49	0.61	0.39	0.28	0.66	0.23	0.41
0.010	10.46	0.23	0.47	0.60	0.38	0.26	0.66	0.23	0.40

how the choice of  $\beta$  influences the performance. **③ Impact of calibration data:** We analyze the effects of calibration dataset selection on compressed model performance in Appendix A.3.4.

**Effectiveness of residual compensation and partial-layer compression.** We present results on C4 [42] and seven zero-shot reasoning and understanding tasks, as shown in Table 4. It can be observed that either applying REC or PLC alone has already improved performance. Combining them leads to further improvements, making the performance gap even more significant. Results of other compression ratios are presented in Table 8.

**Impact of residual compensation factor.** The residual compensation factor  $\beta$  is a hyperparameter in this work, we set  $\beta = 0.05$  across all experiments. To examine its impact on model performance, we compress LLaMA-2-7B under 30% compression ratio while varying the value of  $\beta$ . The results in Table 5 show that the compressed models achieve comparable performance across different  $\beta$  values, demonstrating the robustness of our method with respect to this hyperparameter.

#### 4.4. Compatibility with Quantization

SVD-based LLM compression methods and quantization are two orthogonal techniques. To demonstrate that our method can be integrated with quantization, we adopt GPTQ [12] to quantize LLaMA-2-7B compressed by our method and SVD-LLM. As shown in Table 12, our method integrates seamlessly with quantization, achieving superior performance compared to SVD-LLM.

#### 4.5. Efficiency Results

ERC-SVD not only preserves competitive model performance but also achieves substantial inference speedup on hardware. We evaluate the throughput of compressed models on an NVIDIA A100 GPU, shown in Figure 5. Models compressed by ERC-SVD consistently deliver faster inference than the original model. The speedup grows more pronounced as the batch size grows, indicating that ERC-SVD scales more efficiently under larger workloads.

### 5. Conclusion

In this work, we propose ERC-SVD, a novel post-training SVD-based compression method for LLMs, formulated from an error-controlled perspective. ERC-SVD effectively leverages the residual matrix resulting from SVD truncation to reduce the truncation loss and enhance layer-wise reconstruction accuracy. Furthermore, it selectively compresses only the last few layers of the model, thereby significantly mitigating error propagation. Extensive experiments across various LLM families and benchmark datasets demonstrate that ERC-SVD consistently outperforms existing SVD-based baselines under various settings.

## Acknowledgement

This paper is supported by Young Scientists Fund of the National Natural Science Foundation of China (NSFC) (No. 62506305), Zhejiang Leading Innovative and Entrepreneur Team Introduction Program (No. 2024R01007), Key Research and Development Program of Zhejiang Province (No. 2025C01026), Scientific Research Project of Westlake University (No. WU2025WF003), Chinese Association for Artificial Intelligence (CAAI) & Ant Group Research Fund - AGI Track (No. 2025CAAI-ANT-13). It is also supported by the research funds of the National Talent Program and Hangzhou Municipal Talent Program.

## References

- [1] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- [2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.
- [3] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.
- [4] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Thibaut Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [5] Aixiu Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- [6] 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.
- [7] Ying Sheng, Lianmin Zheng, Binhang Yuan, Zhuohan Li, Max Ryabinin, Beidi Chen, Percy Liang, Christopher Ré, Ion Stoica, and Ce Zhang. Flexgen: High-throughput generative inference of large language models with a single GPU. In *ICML*, 2023.
- [8] Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, et al. A survey on efficient inference for large language models. *arXiv preprint arXiv:2404.14294*, 2024.

- [9] Wenxiao Wang, Wei Chen, Yicong Luo, Yongliu Long, Zhengkai Lin, Liye Zhang, Binbin Lin, Deng Cai, and Xiaofei He. Model compression and efficient inference for large language models: A survey. *arXiv preprint arXiv:2402.09748*, 2024.
- [10] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for modern deep learning research. In *AAAI*, 2020.
- [11] David A. Patterson, Joseph Gonzalez, Urs Hölzle, Quoc V. Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David R. So, Maud Texier, and Jeff Dean. The carbon footprint of machine learning training will plateau, then shrink. *Computer*, 55(7):18–28, 2022.
- [12] 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.
- [13] Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: activation-aware weight quantization for on-device LLM compression and acceleration. In *MLSys*, 2024.
- [14] Wei Huang, Yangdong Liu, Haotong Qin, Ying Li, Shiming Zhang, Xianglong Liu, Michele Magno, and Xiaojuan Qi. Billm: Pushing the limit of post-training quantization for llms. In *ICML*, 2024.
- [15] Zhiteng Li, Xianglong Yan, Tianao Zhang, Haotong Qin, Dong Xie, Jiang Tian, Linghe Kong, Yulun Zhang, Xiaokang Yang, et al. ARB-LLM: alternating refined binarizations for large language models. In *ICLR*, 2025.
- [16] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. In *NeurIPS*, 2023.
- [17] Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *ICML*, 2023.
- [18] Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. In *ICLR*, 2024.
- [19] Shangqian Gao, Chi-Heng Lin, Ting Hua, Zheng Tang, Yilin Shen, Hongxia Jin, and Yen-Chang Hsu. Disp-llm: Dimension-independent structural pruning for large language models. In *NeurIPS*, 2024.
- [20] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *ICLR*, 2024.
- [21] Chuanpeng Yang, Wang Lu, Yao Zhu, Yidong Wang, Qian Chen, Chenlong Gao, Bingjie Yan, and Yiqiang Chen. Survey on knowledge distillation for large language models: Methods, evaluation, and application. *ACM Transactions on Intelligent Systems and Technology*, 2024.
- [22] Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*, 2024.
- [23] Songming Zhang, Xue Zhang, Zengkui Sun, Yufeng Chen, and Jinan Xu. Dual-space knowledge distillation for large language models. In *EMNLP*, 2024.
- [24] Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. Language model compression with weighted low-rank factorization. In *ICLR*, 2022.
- [25] Ayush Kaushal, Tejas Vaidhya, and Irina Rish. LORD: low rank decomposition of monolingual code llms for one-shot compression. *arXiv preprint arXiv:2309.14021*, 2023.

- [26] 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.
- [27] Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. Svd-llm: Truncation-aware singular value decomposition for large language model compression. In *ICLR*, 2025.
- [28] 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.
- [29] Rajarshi Saha, Naomi Sagan, Varun Srivastava, Andrea Goldsmith, and Mert Pilanci. Compressing large language models using low rank and low precision decomposition. In *NeurIPS*, 2024.
- [30] Jingcun Wang, Yu-Guang Chen, Ing-Chao Lin, Bing Li, and Grace Li Zhang. Basis sharing: Cross-layer parameter sharing for large language model compression. In *ICLR*, 2025.
- [31] Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. A survey on model compression for large language models. In *ACL*, 2024.
- [32] Eldar Kurtic, Elias Frantar, and Dan Alistarh. Ziplm: Inference-aware structured pruning of language models. In *NeurIPS*, 2023.
- [33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.
- [34] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *ICLR*, 2020.
- [35] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.
- [36] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [37] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. 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. OPT: open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- [38] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

- [39] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality. *https://vicuna.lmsys.org*, 2(3):6, 2023.
- [40] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *ICLR*, 2017.
- [41] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.
- [42] 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*, 21(140):1–67, 2020.
- [43] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *EMNLP*, 2018.
- [44] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: an adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.
- [45] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *ACL*, 2019.
- [46] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: reasoning about physical commonsense in natural language. In *AAAI*, 2020.
- [47] Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. In *NAACL-HLT*, 2019.
- [48] 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. In *CVPR*, 2023.
- [49] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, 2024.
- [50] Gemma Team. Gemma. 2024. URL <https://www.kaggle.com/m/3301>.
- [51] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023.
- [52] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019.
- [53] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022.
- [54] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.
- [55] Sicheng Feng, Gongfan Fang, Xinyin Ma, and Xinchao Wang. Efficient reasoning models: A survey. *arXiv preprint arXiv:2504.10903*, 2025.
- [56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

- [57] Junhan Zhu, Hesong Wang, Mingluo Su, Zefang Wang, and Huan Wang. Obs-diff: Accurate pruning for diffusion models in one-shot. *arXiv preprint arXiv:2510.06751*, 2025.
- [58] Haolei Bai, Lingcheng Kong, Xueyi Chen, Jianmian Wang, Zhiqiang Tao, and Huan Wang. Dice: Diffusion large language models excel at generating cuda kernels. *arXiv preprint arXiv:2602.11715*, 2026.
- [59] SMTI Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 2024.
- [60] Kejia Zhang, Keda Tao, Jiasheng Tang, and Huan Wang. Poison as cure: Visual noise for mitigating object hallucinations in lvms. *arXiv preprint arXiv:2501.19164*, 2025.

## A. Appendix

Here, we provide further details that are not discussed in the main paper and include extra experimental results. The appendix is structured as follows:

A.1 Pseudocode . . . . .	15
A.2 Layer-wise Error Comparison . . . . .	16
A.3 Additional Results . . . . .	16
A.3.1. Results on Larger-scale LLMs . . . . .	17
A.3.2. Results on Multiple LLM Families . . . . .	17
A.3.3. Additional Ablation Results . . . . .	17
A.3.4. Impact of calibration data . . . . .	17
A.4 Comparison with Smaller-Scale LLM . . . . .	18
A.5 Comparison with Pruning . . . . .	18
A.6 Computation Complexity . . . . .	18
A.7 Demonstration of Generated Contents . . . . .	19
A.8 ERC-SVD on Vision-Language Model . . . . .	19
A.8.1. Quantitative Results . . . . .	19
A.8.2. Qualitative Results . . . . .	20
A.9 Future Work . . . . .	21

### A.1. Pseudocode

Algorithm 2 and Algorithm 3 present the pseudocode for **residual compensation** and **partial-layer compression**, respectively. Algorithm 3 identifies the optimal number of last layers  $k$  to compress, along with their corresponding layer compression ratio  $R_l$ . Specifically, for a model with  $N$  layers, we iterate over candidate values of  $k'$  using a step size  $s$  and compute the associated  $R_l'$  for each. For each candidate configuration, we invoke Algorithm 2 to perform the compression, after which we compute the final-layer error relative to the original model. The configuration yielding the lowest final-layer error is selected as the optimal compression setting.

---

#### Algorithm 2 Pseudocode of Residual Compensation

---

**Input:** Original LLM:  $M$ , weight matrix:  $\mathbf{W}_i \in \mathbb{R}^{m \times n}$ , layer compression ratio:  $R_l$ , residual compensation rank:  $r_r$

**Output:** Compressed weight matrix set  $\text{Set}_{\mathbf{W}'}(i)$  of layer  $i$

- 1:  $CD \leftarrow$  Randomly collect calibration samples from the dataset
  - 2:  $\text{Set}_{\mathcal{S}} \leftarrow \text{WHITENING}(M, CD)$ ,  $\text{Set}_{\mathbf{W}'}(i) \leftarrow \emptyset$  ▷ Initialize sets of weight matrices
  - 3:  $r = (1 - R_l)(m \cdot n) / (m + n)$ ,  $r_i = r - r_r$  ▷ Calculate the intermediate rank
  - 4:  $\mathcal{S}_i \leftarrow \text{Set}_{\mathcal{S}}(i)$  ▷ Extract scaling matrices of current weight matrices
  - 5:  $\mathbf{U}_{i,r_i}, \mathbf{\Sigma}_{i,r_i}, \mathbf{V}_{i,r_i}^T \leftarrow \mathbf{W}_{i,r_i} \mathbf{S}_i^{-1} \leftarrow \text{SVD\_TRUNC}(\mathbf{W}_{i,r_i} \mathbf{S}_i) \mathbf{S}_i^{-1}$  ▷ SVD truncation on weight matrices
  - 6:  $\mathbf{R}_i \leftarrow \text{CAL\_RES}(\mathbf{W}_{i,r_i}, \mathbf{W}_{i,r_i})$  ▷ Calculate residual matrices
  - 7:  $\mathbf{U}_{i,r_r}, \mathbf{\Sigma}_{i,r_r}, \mathbf{V}_{i,r_r}^T \leftarrow \mathbf{R}_i \leftarrow \text{SVD\_TRUNC}(\mathbf{R}_i)$  ▷ SVD truncation on residual matrices
  - 8:  $\hat{\mathbf{U}}_{i,r_r} \leftarrow \text{MUL}(\mathbf{U}_{i,r_r}, \sqrt{\mathbf{\Sigma}_{i,r_r}})$ ,  $\hat{\mathbf{V}}_{i,r_r} \leftarrow \text{MUL}(\sqrt{\mathbf{\Sigma}_{i,r_r}}, \mathbf{V}_{i,r_r}^T)$  ▷ Absorb singular values
  - 9:  $\hat{\mathbf{U}}_{i,r_i} \leftarrow \text{MUL}(\mathbf{U}_{i,r_i}, \sqrt{\mathbf{\Sigma}_{i,r_i}})$ ,  $\hat{\mathbf{V}}_{i,r_i} \leftarrow \text{MUL}(\sqrt{\mathbf{\Sigma}_{i,r_i}}, \mathbf{V}_{i,r_i}^T)$
  - 10:  $\hat{\mathbf{U}}_{i,r} \leftarrow \text{COMB}(\hat{\mathbf{U}}_{i,r_r}, \hat{\mathbf{U}}_{i,r_i})$ ,  $\hat{\mathbf{V}}_{i,r} \leftarrow \text{COMB}(\hat{\mathbf{V}}_{i,r_r}, \hat{\mathbf{V}}_{i,r_i})$  ▷ Combine weight matrices
  - 11:  $\text{Set}_{\mathbf{W}'}(i) \leftarrow (\hat{\mathbf{U}}_{i,r}, \hat{\mathbf{V}}_{i,r}) \cup \text{Set}_{\mathbf{W}'}(i)$  ▷ Append decomposed weight matrices
  - 12: **return**  $\text{Set}_{\mathbf{W}'}(i)$
-

---

**Algorithm 3** Pseudocode of Partial-layer Compression

---

**Input:** Original LLM:  $M$ , number of model layers:  $N$ , overall compression ratio:  $R_o$ , step:  $s$   
**Output:** Number of layers to compress:  $k$ , layer compression ratio:  $R_l$

- 1:  $CD \leftarrow$  Randomly collect calibration samples from the dataset
- 2:  $\text{Set}_{\mathbf{W}} \leftarrow M$  ▷ Extract the set of weight matrices in  $M$
- 3:  $\text{Set}_{k'} \leftarrow \{k' \mid k' = s, 2s, \dots, N - s\}$  ▷ Obtain candidate compressed layer numbers
- 4:  $\text{Lowest\_Err} \leftarrow +\infty$
- 5: **for each**  $k'$  in  $\text{Set}_{k'}$  **do**
- 6:    $R'_l \leftarrow (N \cdot R_o)/k'$  ▷ Calculate corresponding layer compression ratio
- 7:   **if**  $R'_l \geq 0$  **then**
- 8:      $M_{TMP} \leftarrow \text{DEEP\_COPY}(M)$
- 9:     **for**  $i \in [N - k' + 1, N]$  **do**
- 10:        $\mathbf{W}_i \leftarrow \text{Set}_{\mathbf{W}}(i)$  ▷ Initialize weight matrices to compress
- 11:        $\text{Set}_{\mathbf{W}'}(i) \leftarrow \text{RESIDUAL\_COMPENSATION}(M, \mathbf{W}_i, R'_l, r_r)$
- 12:       Replace weights of layer  $i$  in  $M_{TMP}$  with  $\text{Set}_{\mathbf{W}'}(i)$
- 13:     **end for**
- 14:      $\text{Err} \leftarrow \text{CAL\_ERROR}(M, M_{TMP}, CD, \text{Layer}(N))$  ▷ Calculate the last layer output error
- 15:     **if**  $\text{Err} < \text{Lowest\_Err}$  **then**
- 16:        $\text{Lowest\_Err} \leftarrow \text{Err}$  ▷ Search for the lowest layer-wise error
- 17:        $k \leftarrow k'$
- 18:        $R_l \leftarrow R'_l$
- 19:     **end if**
- 20:   **end if**
- 21: **end for**
- 22: **return**  $k, R_l$

---

## A.2. Layer-wise Error Comparison

As discussed in Section 3.2, the same observation also holds here. We conduct experiments on Mistral-7B [38] and Vicuna-7B [39], and the results are presented in Figure 4.

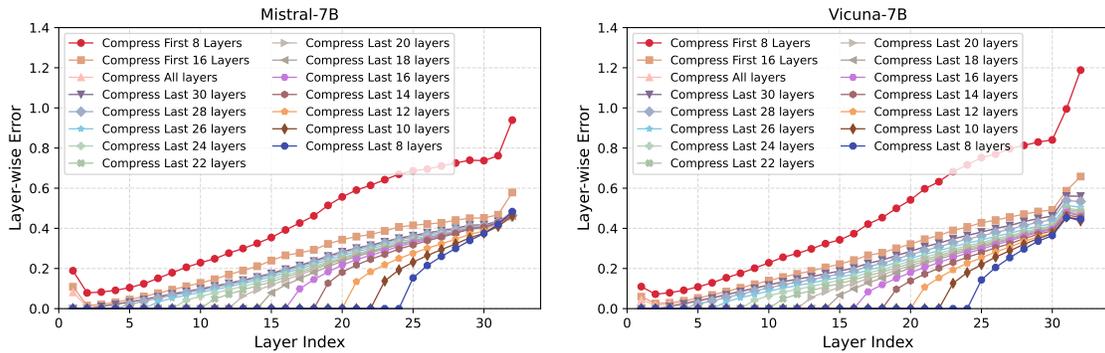


Figure 4: Layer-wise error comparison between the original model, Mistral-7B, and Vicuna-7B compressed by ERC-SVD with different layer selection strategies on WikiText-2. The overall compression ratio is 20%, and all layer selection strategies strictly adhere to the compression constraint.

## A.3. Additional Results

In this section, we present additional results from four perspectives: Section A.3.1 reports zero-shot accuracy on larger-scale LLMs, Section A.3.2 reports zero-shot accuracy across multiple LLM families, Section A.3.3 presents ablation results under compression ratios 40%, 50%, and 60%, and Section A.3.4 demonstrates the impact of calibration data.

Table 6: Zero-shot accuracy results across larger-scale LLMs under 20% compression ratio.

MODEL	METHOD	Openb.↑	ARC_e↑	WinoG.↑	HellaS.↑	ARC_c↑	PIQA↑	MathQA↑	Average↑
LLaMA-13B	SVD-LLM	0.27	0.62	0.66	0.46	0.33	0.72	0.26	0.47
	ERC-SVD	<b>0.28</b>	<b>0.69</b>	<b>0.69</b>	<b>0.54</b>	<b>0.39</b>	<b>0.75</b>	<b>0.26</b>	<b>0.52</b>
LLaMA-30B	SVD-LLM	0.29	0.72	0.73	0.51	0.38	0.74	0.27	0.51
	ERC-SVD	<b>0.30</b>	<b>0.73</b>	<b>0.73</b>	<b>0.59</b>	<b>0.43</b>	<b>0.78</b>	<b>0.29</b>	<b>0.55</b>
LLaMA-2-13B	SVD-LLM	0.27	0.63	0.66	0.44	0.31	0.71	0.26	0.47
	ERC-SVD	<b>0.31</b>	<b>0.67</b>	<b>0.68</b>	<b>0.47</b>	<b>0.38</b>	<b>0.74</b>	<b>0.27</b>	<b>0.50</b>
OPT-13B	SVD-LLM	0.24	0.62	0.64	<b>0.47</b>	0.30	0.72	0.24	0.46
	ERC-SVD	<b>0.24</b>	<b>0.63</b>	<b>0.65</b>	0.45	<b>0.30</b>	<b>0.73</b>	<b>0.24</b>	<b>0.46</b>
OPT-30B	SVD-LLM	<b>0.29</b>	0.67	0.65	<b>0.51</b>	0.33	0.75	0.25	0.49
	ERC-SVD	0.28	<b>0.68</b>	<b>0.67</b>	0.50	<b>0.34</b>	<b>0.76</b>	<b>0.25</b>	<b>0.49</b>

Table 7: Performance of Qwen-3-8B under 20% to 60% compression ratios (“RATIO”).

RATIO	METHOD	WikiText-2↓	PTB↓	C4↓	Openb.↑	ARC_e↑	WinoG.↑	HellaS.↑	ARC_c↑	PIQA↑	MathQA↑	Average↑
	Original	9.71	15.43	15.52	0.31	0.83	0.68	0.57	0.56	0.77	0.49	0.60
20%	SVD-LLM	37.52	<b>40.73</b>	47.25	<b>0.20</b>	0.50	0.55	0.35	0.24	0.64	0.22	0.39
	ERC-SVD	<b>35.11</b>	42.77	<b>43.10</b>	0.19	<b>0.54</b>	<b>0.58</b>	<b>0.35</b>	<b>0.25</b>	<b>0.66</b>	<b>0.24</b>	<b>0.40</b>
30%	SVD-LLM	72.33	67.68	86.56	0.15	<b>0.42</b>	0.53	0.30	0.19	0.60	0.21	0.34
	ERC-SVD	<b>66.89</b>	<b>71.48</b>	<b>82.28</b>	<b>0.16</b>	0.41	<b>0.53</b>	<b>0.30</b>	<b>0.20</b>	<b>0.61</b>	<b>0.22</b>	<b>0.35</b>
40%	SVD-LLM	143.28	122.55	162.99	0.12	0.34	0.51	0.28	0.16	<b>0.58</b>	0.22	0.32
	ERC-SVD	<b>135.66</b>	<b>113.34</b>	<b>154.32</b>	<b>0.12</b>	<b>0.35</b>	<b>0.52</b>	<b>0.29</b>	<b>0.17</b>	0.57	<b>0.22</b>	<b>0.32</b>
50%	SVD-LLM	274.04	226.30	308.11	0.12	0.30	0.50	0.27	0.18	0.55	0.20	0.30
	ERC-SVD	<b>247.57</b>	<b>170.82</b>	<b>294.00</b>	<b>0.12</b>	<b>0.30</b>	<b>0.50</b>	<b>0.28</b>	<b>0.18</b>	<b>0.56</b>	<b>0.22</b>	<b>0.31</b>
60%	SVD-LLM	462.53	338.39	528.22	0.11	0.27	<b>0.48</b>	0.26	0.18	<b>0.54</b>	0.20	0.29
	ERC-SVD	<b>460.73</b>	<b>287.19</b>	<b>522.29</b>	<b>0.11</b>	<b>0.27</b>	0.46	<b>0.27</b>	<b>0.19</b>	0.53	<b>0.20</b>	<b>0.29</b>

### A.3.1. Results on Larger-scale LLMs

In Table 6, we present zero-shot accuracy results for LLaMA-13B [4], LLaMA-30B [4], LLaMA-2-13B [36], OPT-13B [37], and OPT-30B [37]. Across these evaluation datasets, ERC-SVD consistently outperforms SVD-LLM, with only one minor exception.

### A.3.2. Results on Multiple LLM Families

We also evaluate the zero-shot accuracy of OPT-6.7B [37], Mistral-7B [38], and Vicuna-7B [39]. Results are shown in Table 9, and our method consistently outperforms SVD-LLM across these diverse LLM architectures. In addition, we conduct experiments on Qwen-3-8B [6], with perplexity and zero-shot accuracy results reported in Table 7.

### A.3.3. Additional Ablation Results

The ablation results for LLaMA-2-7B [36] under 40%, 50%, and 60% compression ratios are presented in Table 8. A similar trend can also be observed here: incorporating both REC and PLC leads to a substantial reduction in perplexity across all settings.

### A.3.4. Impact of calibration data

We study the impact of calibration data from two aspects: the number of calibration samples and the choice of calibration datasets. Figure 6 shows the evaluation results under 30% compression ratio. Performance improves slightly with more calibration samples, though the gains remain modest, indicating that ERC-SVD is robust even with limited calibration data. Moreover, Table 11 reports results for different calibration datasets. We also evaluate a mixed dataset, constructed by equally combining the three datasets, to assess its effect on model performance.

Table 8: Ablation results of residual compensation and partial-layer compression on LLaMA-2-7B [36] under 40%, 50%, and 60% compression ratios.

RATIO	METHOD	REC	PLC	C4↓	Openb.↑	ARC_e↑	WinoG.↑	HellaS.↑	ARC_c↑	PIQA↑	MathQA↑	Average↑
40%	ASVD	-	-	NaN	0.15	0.25	0.50	0.26	0.22	0.52	0.18	0.30
	SVD-LLM	-	-	61.96	0.16	0.35	0.55	0.30	0.20	0.57	0.23	0.34
	ERC-SVD	✓	✗	54.19	0.17	0.37	0.54	0.32	0.21	0.57	0.23	0.34
		✗	✓	45.13	0.18	0.40	0.55	0.33	0.24	0.61	0.21	0.36
		✓	✓	<b>43.19</b>	0.20	0.43	0.57	0.35	0.24	0.63	0.23	<b>0.38</b>
50%	ASVD	-	-	NaN	0.13	0.26	0.50	0.25	0.23	0.50	0.20	0.30
	SVD-LLM	-	-	129.71	0.14	0.30	0.50	0.28	0.20	0.54	0.23	0.31
	ERC-SVD	✓	✗	126.61	0.14	0.30	0.53	0.29	0.21	0.55	0.24	0.32
		✗	✓	117.79	0.13	0.33	0.54	0.29	0.22	0.58	0.22	0.33
		✓	✓	<b>100.34</b>	0.14	0.35	0.55	0.31	0.22	0.59	0.22	<b>0.34</b>
60%	ASVD	-	-	NaN	0.15	0.25	0.50	0.25	0.23	0.52	0.12	0.29
	SVD-LLM	-	-	263.02	0.14	0.26	0.50	0.27	0.20	0.53	0.21	0.30
	ERC-SVD	✓	✗	256.38	0.13	0.26	0.49	0.27	0.20	0.53	0.23	0.30
		✗	✓	260.01	0.14	0.29	0.50	0.26	0.18	0.53	0.21	0.30
		✓	✓	<b>255.70</b>	0.13	0.29	0.52	0.28	0.21	0.55	0.22	<b>0.31</b>

Table 9: Zero-shot accuracy (↑) for seven common sense reasoning datasets on OPT-6.7B, Mistral-7B, and Vicuna-7B under 30% compression ratio.

MODEL	METHOD	Openb.↑	ARC_e↑	WinoG.↑	HellaS.↑	ARC_c↑	PIQA↑	MathQA↑	Average↑
OPT-6.7B	SVD-LLM	0.27	0.40	0.49	0.33	0.22	0.50	0.20	0.34
	ERC-SVD	0.22	<b>0.56</b>	<b>0.61</b>	<b>0.40</b>	<b>0.25</b>	<b>0.69</b>	<b>0.24</b>	<b>0.42</b>
Mistral-7B	SVD-LLM	0.13	0.44	0.53	0.30	0.20	0.62	0.20	0.34
	ERC-SVD	<b>0.14</b>	<b>0.46</b>	<b>0.58</b>	<b>0.33</b>	<b>0.22</b>	<b>0.63</b>	<b>0.24</b>	<b>0.37</b>
Vicuna-7B	SVD-LLM	0.22	0.50	0.56	0.35	0.26	0.63	0.21	0.39
	ERC-SVD	<b>0.23</b>	<b>0.51</b>	<b>0.60</b>	<b>0.39</b>	<b>0.30</b>	<b>0.67</b>	<b>0.22</b>	<b>0.42</b>

Table 10: Results on Wikitext-2 under different compression ratios ("RATIO") of LLaMA-2-7B.

RATIO	LLM-Pruner	ERC-SVD
20%	8.09	7.63
30%	12.59	10.32
40%	20.36	14.17
50%	40.97	24.26
60%	114.23	58.88

#### A.4. Comparison with Smaller-Scale LLM

To further assess the effectiveness of ERC-SVD under relatively high compression ratios, we conduct a comparative evaluation of 50% and 60% compressed LLaMA-2-7B against Gemma-2-2B [50]. Table 13 presents the performance comparison between the compressed LLaMA-2-7B models and Gemma-2-2B. Under 50% compression ratio (approximately 3.5B parameters), LLaMA-2-7B consistently outperforms Gemma-2-2B across most benchmarks. Even under a more aggressive 60% compression ratio (around 2.8B parameters), the compressed model maintains competitive performance, remaining comparable to Gemma-2-2B.

#### A.5. Comparison with Pruning

Table 10 shows the performance of LLaMA-2-7B compressed by LLM-Pruner [16] and ERC-SVD under different compression ratios on the WikiText-2 dataset. It can be observed that ERC-SVD consistently outperforms the pruning method, achieving a perplexity of 58.88 under 60% compression ratio, compared to 114.23 for LLM-Pruner.

#### A.6. Computation Complexity

ERC-SVD decomposes the original weight matrix  $\mathbf{W} \in \mathbb{R}^{m \times n}$  into two low-rank matrices:  $\hat{\mathbf{U}}_r \in \mathbb{R}^{m \times r}$  and  $\hat{\mathbf{V}}_r \in \mathbb{R}^{r \times n}$ . The layer compression ratio  $R_l$  is computed as  $R_l = 1 - \frac{(m+n)r}{mn}$ . Under a fixed overall compression ratio  $R_o$ , the relationship between  $R_l$  and  $R_o$  is given by  $R_l = \frac{NR_o}{k}$ , where  $N$  is the total number of layers and  $k$  denotes the number of last layers to be compressed.

Given an input activation  $\mathbf{X} \in \mathbb{R}^{n \times m}$ , the original output is computed as  $\mathbf{Y} = \mathbf{W}\mathbf{X}$ . In the compressed model layer, an intermediate state is first computed as  $\mathbf{I} = \hat{\mathbf{V}}_r\mathbf{X}$ , followed by  $\mathbf{Y} = \hat{\mathbf{U}}_r\mathbf{I}$ . The computational complexity of the original model is  $N \cdot \mathcal{O}(m^2n)$ . For the compressed model, the first  $(N - k)$  layers remain uncompressed and retain a complexity of  $(N - k) \cdot \mathcal{O}(m^2n)$ , while the

Table 11: Perplexity ( $\downarrow$ ) and zero-shot average accuracy ( $\uparrow$ ) of LLaMA-2-7B 20% compressed by ERC-SVD with different calibration datasets.

CAL. DATA	WikiText-2 $\downarrow$	PTB $\downarrow$	C4 $\downarrow$	Avg. $\uparrow$
WikiText-2	<b>7.63</b>	45.37	14.73	<b>0.48</b>
PTB	9.71	<b>29.43</b>	13.29	0.46
C4	9.44	37.67	<b>11.49</b>	0.47
Mix	10.13	36.53	13.13	0.45

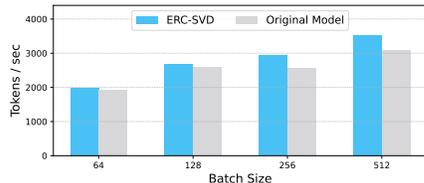


Figure 5: Throughput of LLaMA-7B and its 40% compressed versions. The sequence length is 32.

Table 12: Perplexity ( $\downarrow$ ) of LLaMA-2-7B compressed by ERC-SVD and SVD-LLM, followed by quantization with GPTQ-8bit [12]. Blue arrows within parentheses highlight the relative improvement.

RATIO	METHOD	QUANTIZATION	WikiText-2 $\downarrow$	PTB $\downarrow$	C4 $\downarrow$
-	Original	GPTQ-8bit	5.47	26.79	7.28
30%	SVD-LLM	GPTQ-8bit	10.67	336.30	34.93
	<b>ERC-SVD</b>	<b>GPTQ-8bit</b>	<b>10.34</b> ( $\downarrow$ 3%)	<b>79.05</b> ( $\downarrow$ 76%)	<b>25.02</b> ( $\downarrow$ 28%)

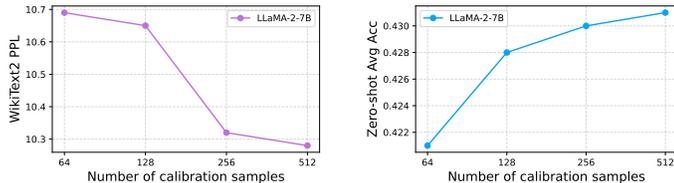


Figure 6: Impact of the number of calibration data samples on LLaMA-2-7B under 30% compression ratio. (Left) Perplexity ( $\downarrow$ ). (Right) Average accuracy ( $\uparrow$ ).

compressed  $k$  layers incur a cost of  $k \cdot \mathcal{O}(m^2r + rnm)$ . Thus, the total complexity becomes:

$$(N - k) \cdot \mathcal{O}(m^2n) + k \cdot \mathcal{O}(m^2r + rnm). \quad (15)$$

And the rank  $r$  is given by:

$$r = \frac{mn(1 - R_l)}{m + n} = \frac{mn(k - NR_o)}{k(m + n)}. \quad (16)$$

Substituting the expression for  $r$  into Equation 15, we obtain the simplified total complexity:

$$N(1 - R_o) \cdot \mathcal{O}(m^2n) \quad (17)$$

Compared to the original computation complexity  $N \cdot \mathcal{O}(m^2n)$ . This indicates that the total computation cost is reduced proportionally to the overall compression ratio. For example, if  $R_o = 40\%$ , the compressed model requires only 60% of the original computational cost.

## A.7. Demonstration of Generated Contents

Table 15 shows the generation contents of models compressed by ERC-SVD compared to the original model under zero-shot conditions. The results indicate that across various input questions, compressed models consistently produce fluent, coherent, and highly relevant responses. Even at 40% compression ratio, the compressed model can also deliver accurate and relevant answers to common sense questions. These results highlight the effectiveness of ERC-SVD.

## A.8. ERC-SVD on Vision-Language Model

In this section, we apply ERC-SVD to the language part of vision-language models (VLMs) such as LLaVA [49], and evaluate both quantitative and qualitative performance.

### A.8.1. Quantitative Results

We compare the performance of LLaVA-1.5-7B [49] compressed using SVD-LLM [27] and ERC-SVD under 20% compression ratio. We report results on several benchmarks: POPE [51], TextVQA [52], and ScienceQA [53]. The results, shown in Table 14, indicate that ERC-SVD consistently outperforms SVD-LLM across all benchmarks. Notably, it achieves substantial relative improvements, 66% on TextVQA and 40% on ScienceQA. Moreover, on the POPE-random and POPE-popular subsets, the model compressed by ERC-SVD even surpasses the original LLaVA-1.5-7B.

Table 13: Performance comparison of Gemma-2-2B and LLaMA-2-7B compressed by ERC-SVD under 50% and 60% compression ratios (“RATIO”).

MODEL	RATIO	#Params	WikiText-2↓	PTB↓	C4↓	Openb.↑	ARC_e↑	WinoG.↑	HellaS.↑	ARC_c↑	PIQA↑	MathQA↑	Average↑
Gemma-2-2B	-	3B	67.78	374.60	70.78	0.13	0.39	0.51	0.31	0.19	0.60	0.23	0.33
LLaMA-2-7B	50%	3.5B	24.26	286.24	100.34	0.14	0.35	0.55	0.31	0.22	0.59	0.22	0.34
	60%	2.8B	68.59	991.48	255.70	0.13	0.29	0.52	0.28	0.21	0.55	0.22	0.31

Table 14: Performance (measured by accuracy (↑)) of original LLaVA-1.5-7B, and its 20% compressed versions by SVD-LLM and ERC-SVD on VLM benchmarks. The best results are marked in **bold**.

COMPRESSION RATIO	METHOD	POPE-random↑	POPE-popular↑	POPE-adversarial↑	TextVQA↑	ScienceQA↑
-	Original	88.2	87.3	85.1	58.17	70.15
20%	SVD-LLM	82.5	83.2	77.8	30.68	49.54
	ERC-SVD	<b>90.2 (↑9%)</b>	<b>87.7 (↑5%)</b>	<b>83.1 (↑7%)</b>	<b>50.86 (↑66%)</b>	<b>69.54 (↑40%)</b>

### A.8.2. Qualitative Results

As shown in Figure 7, LLaVA-1.5-7B compressed with ERC-SVD under 20% compression ratio is still able to produce image captions that are faithful to the visual content. This indicates that the model retains its ability to interpret and describe images accurately, despite parameter reduction.

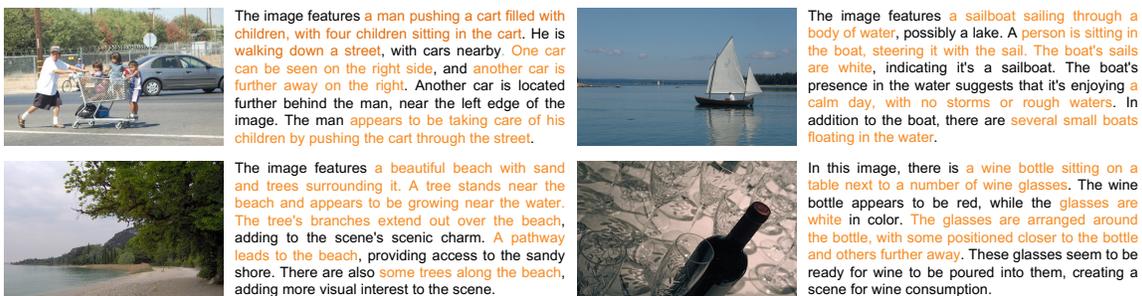


Figure 7: Image captioning results of LLaVA-1.5-7B compressed by ERC-SVD under 20% compression ratio. Captions that accurately describe the image content are highlighted in orange.

Figure 8 presents the visual question answering results of LLaVA-1.5-7B compressed by ERC-SVD. The model retains strong performance across these question types, indicating that its multi-modal reasoning capability remains intact despite the compression.

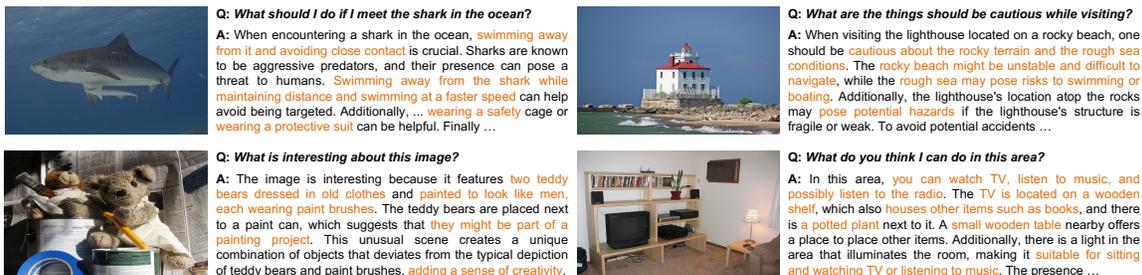


Figure 8: Visual question answering outputs generated by LLaVA-1.5-7B compressed using ERC-SVD under 20% compression ratio. Questions (Q) and model answers (A) are provided, correct answers are highlighted in orange to emphasize answer quality retention.

## A.9. Future Work

While this work focuses on general LLM compression, the low-rank structures identified via SVD hold significant potential for reasoning LLMs [6, 54, 55]. Beyond autoregressive language tasks, the principles of SVD-based compression are readily transferable to diffusion frameworks [56–58]. In related tasks, the Transformer-based denoisers often exhibit extreme parameter redundancy. By leveraging ERC-SVD to decompose attention layers, maybe we can accelerate the iterative sampling process, making real-time deployment of generative models on edge devices more feasible without sacrificing structural fidelity. Furthermore, the intersection of model compression and hallucination mitigation [59, 60] presents an intriguing research frontier. We hypothesize that low-rank approximations could potentially filter out the noise that contributes to factually incorrect generations, a direction we intend to explore in future work.

Table 15: Generated contents of LLaMA-7B [4] compressed by ERC-SVD.

QUESTION	COMP. RATIO	ANSWER
What is the responsibility of the AI assistant?	-	The AI assistant is responsible for understanding the context and taking appropriate action. The AI assistant is responsible for understanding the intent, context, environment and taking appropriate actions. The AI assistant is responsible for interpreting.
	20%	The AI assistant should be responsible for providing an answer to questions. This means that when asked a question, you should do your best to answer it and not give any information other than what is requested. You should also consider the tone and intent.
	30%	The AI Assistant's job is to provide assistance to those who wish for assistance in their lives. It has no duty or obligation to assist those who do not desire assistance, nor does it have any duty to prevent.
	40%	The AI assistant does not have a specific role. The AI assistant will provide instructions to follow a path, and will give directions as an instruction to follow the path.
What is the universe? Explain this concept.	-	The Universe is everything that exists. It includes all space, time, matter and energy. It also includes everything which has happened or will happen. The universe began about 13.7 billion years ago (the Big Bang) and will.
	20%	The universe is the whole of space and time. It consists of all things in space and all things that happen, or have happened. There are many theories on how it began; there were theories that it had no beginning at all.
	30%	The universe consists of everything in it, including matter and objects, space, the cosmos or outer space, and the galaxy system, as well as other galaxies that make up the Universe. The universe has its own life cycle.
	40%	In the universe, life is very diverse and complex. The universe is made up of many different objects. These objects are divided into groups. Each group consists of different objects that have specific properties. This group contains the universe, which is large.
Why do cars need fuel to run?	-	Fuel is the lifeblood of a car. Without it, there's nothing to power the pistons and no way to drive. A car needs a constant supply of fuel.
	20%	The fuel in the tank is burned to produce heat, which produces steam that drives a turbine, which turns a crankshaft. The crankshaft turns the wheels of the car, propelling it forward.
	30%	The fuel is the fuel in a car and it burns and goes up and up. It's important because it burns the gasoline, which makes it go faster. It's important because you can get energy out.
	40%	The fuel gives energy for the cars. The gasoline is used by a motor, which produces heat and air, heating up the engine. This allows the engines to travel faster, allowing them to reach higher speed.
Why do people feel hungry after exercise?	-	The answer is because the body burns more calories when you are exercising. You may not realize that the body uses up energy to digest food and absorb nutrients from it.
	20%	Because the body is more efficient at converting food into energy when inactive, so when active it needs to convert food into energy faster. This is why people often get hungry after exercise.
	30%	The reason people feel hungry after an activity is because they burn energy and energy is metabolised in the body.
	40%	The reason the body is hungry after a long athletic activity was discovered by scientists in 1976.