NMP-SVD: RETHINKING DENSE PROJECTIONS VIA N:M-STRUCTURED LOW-RANK COMPRESSION FOR LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing large language model (LLM) compression methods based on Singular Value Decomposition (SVD) mainly focus on weighted reconstruction, singular-value truncation, and rank allocation, while leaving the representation of retained projection directions largely unexplored. Most methods preserve these projections as dense factors. We revisit this assumption through controlled perturbation experiments and find that low-rank representations can tolerate substantial directional approximation, with expanded approximate projections often outperforming smaller exact dense representations. Motivated by this observation, we propose NMP-SVD (N:M Projection SVD), a plug-and-play structured-sparsity enhancement for SVD compressors. NMP-SVD optimizes both sparse low-rank factors under the original weighted reconstruction objective using dual-factor N:M-OBS calibration, expands projection directions from the sparse residual, and applies block-wise backfitting to correct sequential expansion errors. A per-layer selector retains dense realizations for sparsity-sensitive layers. NMP-SVD preserves the weighting and rank-allocation strategies of the underlying compressor, enabling compatibility with diverse SVD frameworks. Experiments across multiple models and datasets demonstrate improved quality–compression trade-offs, higher prefill efficiency, and reduced memory usage. Code is available at `https://anonymous.4open.science/r/NMP-SVD-279D`.

1 INTRODUCTION

Large language models (LLMs) (Touvron et al., 2023; Dubey et al., 2024; Yang et al., 2025) have achieved remarkable capability in language understanding, generation, reasoning, and code, but their rapidly growing parameter counts impose substantial storage, memory-bandwidth, and inference-compute costs. Reducing deployment costs while preserving model capabilities as much as possible has become a key challenge in large-model systems research. Quantization, pruning, and low-rank factorization represent three major directions (Zhu et al., 2024; Xu & McAuley, 2023). Among them, SVD-based low-rank compression replaces a large linear weight matrix with the product of two smaller factors, straightforward compatibility with dense linear algebra, and natural complementarity with other compression techniques.

Recent advances in SVD-based LLM compression have primarily improved reconstruction objectives through activation-aware(Yuan et al., 2023) and Fisher-aware metrics (Thoma et al., 2026), refined truncation and weight-reconstruction procedures (Wang et al., 2025a), and developed adaptive rank-allocation (Abbasi et al., 2026) strategies across layers. Despite their strong performance, existing SVD-based LLM compression methods typically represent low-rank factors as dense projections. Truncated SVD provides an optimal rank-r approximation under the Frobenius norm. **However, this optimality does not guarantee that dense factors are optimal for LLM compression under a fixed deployment budget.** Yet prior work has offered limited fine-grained analysis of how these projections should be represented to balance model capability and deployment cost.

We study this question through controlled perturbation experiments that isolate projection representation from any particular sparse solver, as shown in Figure 2. Across LLaMA-2, LLaMA-3, and Qwen3, an expanded set of moderately approximated projection directions can outperform the ex-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

act Dense-r representation, indicating that low-rank compression can tolerate substantial deviations from individually exact projections. This tolerance varies markedly across layers and linear-layer types. However, when instantiated with actual 2:4 sparsity, naive post-hoc masking cannot reliably preserve this advantage, whereas low-rank-aware structured optimization is substantially more stable. These observations shift the question from whether dense low-rank factors can be sparsified to how projection representations should be optimized directly under $N{:}M$ constraints.

Based on these findings, we introduce NMP-SVD ($N{:}M$ Projection SVD), a plug-and-play projection-realization framework for SVD-based LLM compression. NMP-SVD preserves the weighting mechanism and rank allocation of the underlying compressor while changing how its low-rank projections are realized. It performs alternating dual-factor $N{:}M$ optimization under the original weighted reconstruction objective, expands projection directions from the residual of the actual sparse solution, and jointly refines the resulting factors. To account for layer-wise heterogeneity in sparsity tolerance, NMP-SVD further selects between dense and sparse realizations at the level of individual linear layers. As illustrated in Figure 1, this structured realization can move existing SVD compressors toward a better quality–efficiency trade-off while also reducing the storage-equivalent parameter count and runtime memory footprint. Our main contributions are summarized as follows:

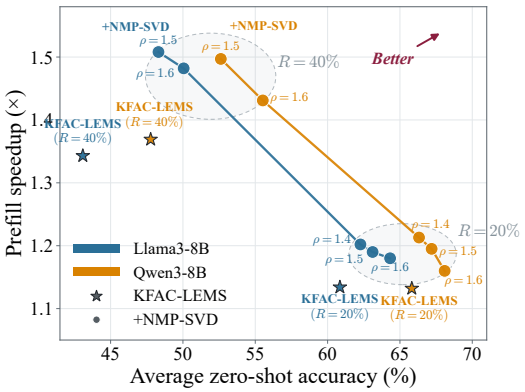


Figure 1: Quality–efficiency trade-off of KFAC-LEMS with NMP-SVD. R denotes the compression ratio and ρ the expansion ratio. NMP-SVD improves average zero-shot accuracy and prefill speed across the evaluated operating points.

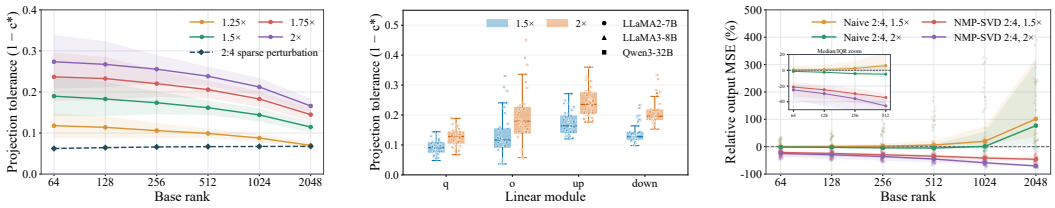
- We identify *projection realization* as an overlooked design dimension in SVD-based LLM compression, and show that broader but approximate projection representations can outperform narrower exact Dense-r counterparts under constrained deployment budgets.
- We formulate N:M-constrained projection realization directly in the original weighted reconstruction objective and introduce NMP-SVD, a factor-aware solver that couples dual-factor OBS calibration with sparse-residual projection expansion.
- NMP-SVD is plug-and-play with SVD weighting and rank-allocation schemes, improving quality–compression trade-offs while reducing memory and accelerating prefill inference.

2 Related Work

Existing work combines sparsity and low-rank modeling, sush as 3BASiL (Makni et al., 2026) and SLiM (Mozaffari et al., 2024), mainly through sparse-plus-low-rank decomposition or low-rank residual compensation. Sparse SVD has also been studied (Lee et al., 2010), typically with unstructured sparsity; N:M low-rank factors and their impact on LLM compression remain underexplored.

2.1 SVD-based Low-Rank Compression for LLMs

SVD-based LLM compression replaces large weight matrices with low-rank factors, but vanilla truncation often degrades performance because different directions contribute unequally to model outputs and task loss. ASVD (Yuan et al., 2023) incorporates activation statistics and sensitivity-aware rank search, while SVD-LLM (Wang et al., 2025c) introduces truncation-aware whitening and sequential parameter updates. SVD-LLM V2 (Wang et al., 2025b) further improves truncation and heterogeneous rank allocation. Recent work extends SVD compression toward loss sensitivity and global optimization. Dobi-SVD (Wang et al., 2025a) jointly optimizes truncation positions and reconstruction; Zero-Sum SVD (Abbasi et al., 2026) performs loss-aware global singular-component selection; and KFAC-SVD with LEMS (Thoma et al., 2026) combines two-sided Fisher-weighted decomposition with global rank allocation. Overall, prior work mainly focuses on weighting, truncation, reconstruction, and rank allocation, while the low-rank factors remain predominantly dense.



(a) Tolerance across ranks. (b) Model/layer-wise heterogeneity. (c) N:M-aware sparse realization

Figure 2: **Rethinking dense low-rank projections.**

2.2 N:M Semi-Structured Sparsity

N:M semi-structured sparsity retains N nonzero weights within every group of M consecutive parameters, providing a regular pattern that combines pruning flexibility with hardware-friendly structure. In particular, 2:4 sparsity is supported by modern sparse tensor cores, making it attractive for accelerating LLM inference. Training-free pruning methods mainly improve sparse mask selection and compensate for pruning-induced reconstruction error. SparseGPT (Frantar & Alistarh, 2023) exploits second-order information and sequential error compensation, while Wanda (Sun et al., 2024) combines weight magnitude with activation statistics to identify important weights. Recent studies further improve N:M sparsification through cross-layer curvature modeling (Bai et al., 2024), rotation-based transformations (Gu et al., 2026), permutation strategies (Zou et al., 2026), and learnable sparse masks (Liu et al., 2025). Collectively, these methods show that effective N:M sparsification depends on reliable importance estimation and appropriate error compensation.

3 Is Dense Projection Necessary?

We first ask a fundamental question: **Must SVD projections in LLM compression remain fully dense?** To isolate projection representation from any specific sparsification algorithm, we conduct controlled directional perturbation experiments. Let D_r denote the exact Dense-r approximation and expand it to rank ρr. We perturb the expanded projection directions to a prescribed cosine similarity c and measure the resulting layer-output MSE. We define the critical similarity $c^*(r, \rho)$ as the lowest similarity at which the perturbed representation remains no worse than Dense-r, and define the projection tolerance as $\tau(r, \rho) = 1 - c^*(r, \rho)$.

To estimate c^* without assuming global monotonicity of the reconstruction error with respect to c, we first scan the full similarity interval on a uniform grid to locate the lowest feasible transition, and then refine the corresponding local interval by bisection. A larger τ therefore indicates greater tolerance to directional approximation. Perturbations are sampled from the orthogonal residual subspace to preserve vector norms and avoid uncontrolled correlation. A simple spectral approximation provides intuition for this trade-off. If the captured spectral energy grows approximately with rank, the admissible perturbation scales as

$$\tau_{\text{approx}}(\rho) \approx \frac{\rho - 1}{2\rho}, \tag{1}$$

which increases from approximately 0.10 to 0.25 as ρ grows from 1.25 to 2. The exact tolerance remains spectrum-dependent; a derivation is provided in Appendix A.1.

Figure 2(a) confirms this trend empirically: increasing ρ consistently enlarges the admissible directional deviation across base ranks. Notably, magnitude-based 2:4 sparsification introduces an average cosine distance of only 0.062–0.068, while using approximately 75% as many nonzeros as Dense-r at $\rho = 1.5$. **This perturbation lies within the observed tolerance range for many configurations, suggesting that exact dense projection directions are not always necessary.**

Figure 2(b) further shows that this trend generalizes across LLaMA-2-7B, LLaMA3-8B, and Qwen3-32B, but not uniformly across layers and projection types. The model-level median tolerance increases from approximately 0.112–0.149 at 1.5× expansion to 0.169–0.218 at 2×. While increased projection coverage generally improves tolerance, some layers remain substantially more sparsity-sensitive, motivating per-linear-layer adaptive dense–sparse realization.

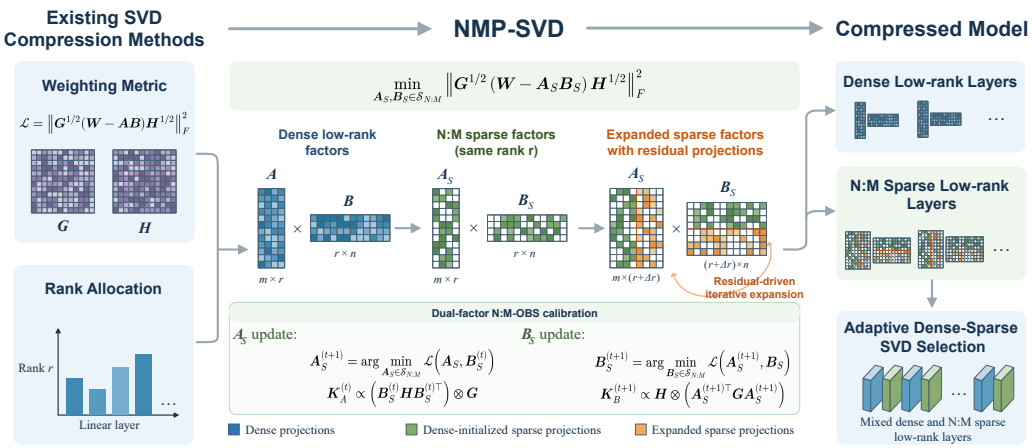


Figure 3: Overview of NMP-SVD. Given the base weighting metric and base rank by an existing SVD compressor, NMP-SVD performs dual-factor $N{:}M$-OBS calibration, sparse-residual expansion, block-wise backfitting, and adaptive dense–sparse selection.

Finally, Figure 2(c) replaces synthetic perturbations with actual 2:4 sparse factors. Naive post-hoc sparsification matches or improves upon Dense-r in only about 19%–25% of configurations at $1.5\times$ expansion and 64%–72% at $2\times$, with substantial failures remaining. In contrast, the $N{:}M$-aware low-rank solver (our NMP-SVD) achieves substantially lower and more stable relative output MSE. Thus, projection tolerance alone is insufficient: exploiting it effectively requires optimizing the low-rank representation directly under structured sparsity constraints.

4 NMP-SVD

4.1 Problem Formulation

Consider a linear-layer weight matrix $\boldsymbol{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. A rank-$r$ approximation is written as $\boldsymbol{W} \approx \boldsymbol{AB}$, where $\boldsymbol{A} \in \mathbb{R}^{d_{\text{out}} \times r}$ and $\boldsymbol{B} \in \mathbb{R}^{r \times d_{\text{in}}}$. To cover activation-aware and Fisher-aware SVD compressors, we use the two-sided weighted reconstruction objective:

$$\mathcal{L}(\boldsymbol{A}, \boldsymbol{B}) = \left\| \boldsymbol{G}^{1/2}(\boldsymbol{W} - \boldsymbol{AB})\boldsymbol{H}^{1/2} \right\|_F^2, \tag{2}$$

where $\boldsymbol{H} \succeq 0$ and $\boldsymbol{G} \succeq 0$ denote the input- and output-side importance metrics, respectively.

Existing weighted SVD compressors typically absorb the weighting matrices into a transformed representation, e.g., $\overline{\boldsymbol{W}} = \boldsymbol{G}^{1/2}\boldsymbol{W}\boldsymbol{H}^{1/2}$, and perform low-rank decomposition in this weighted space. The resulting factors are then mapped back through $\boldsymbol{G}^{-1/2}$ and $\boldsymbol{H}^{-1/2}$. Although this transformation is suitable for dense low-rank decomposition, directly imposing $N{:}M$ sparsity in the weighted space is problematic because the inverse transformations generally destroy the prescribed sparsity pattern. NMP-SVD therefore keeps $\boldsymbol{G}$ and $\boldsymbol{H}$ explicit in the reconstruction objective and imposes the $N{:}M$ constraints directly on the final deployable factors. Specifically, it solves

$$\begin{aligned} \min_{\boldsymbol{A}_S, \boldsymbol{B}_S} \quad & \left\| \boldsymbol{G}^{1/2}(\boldsymbol{W} - \boldsymbol{A}_S\boldsymbol{B}_S)\boldsymbol{H}^{1/2} \right\|_F^2, \\ \text{s.t.} \quad & \boldsymbol{A}_S, \boldsymbol{B}_S \in \mathbb{S}_{N:M}, \qquad r' \geq r, \end{aligned} \tag{3}$$

where $\boldsymbol{A}_S \in \mathbb{R}^{d_{\text{out}} \times r'}$, $\boldsymbol{B}_S \in \mathbb{R}^{r' \times d_{\text{in}}}$, and r' denotes the factor width. This formulation preserves the weighting criterion of the underlying SVD compressor while enforcing structured sparsity directly on the factors used at inference time. The framework of our NMP-SVD is shown in Figure 3.

4.2 Dual-Factor $N{:}M$-OBS Calibration

The constrained problem in Eq. 3 naturally aligns with the OBS formulation. Once one low-rank factor is fixed, the objective becomes a quadratic weighted reconstruction problem with respect to the

other factor, while imposing $N{:}M$ sparsity amounts to removing a structured subset of parameters under this quadratic metric. OBS therefore provides a principled way to evaluate the reconstruction loss induced by sparsification and to compensate for the removed parameters.

Because the two factors are multiplicatively coupled, NMP-SVD adopts a Gauss–Seidel-style alternating optimization. At iteration t, the left factor is first optimized conditioned on the current right factor $\boldsymbol{B}_S^{(t)}$; the updated $\boldsymbol{A}_S^{(t+1)}$ is then immediately used to optimize the right factor:

$$\boldsymbol{A}_S^{(t+1)} = \arg\min_{\boldsymbol{A}_S \in \mathbb{S}_{N:M}} \mathcal{L}\Big(\boldsymbol{A}_S, \boldsymbol{B}_S^{(t)}\Big), \qquad \boldsymbol{B}_S^{(t+1)} = \arg\min_{\boldsymbol{B}_S \in \mathbb{S}_{N:M}} \mathcal{L}\Big(\boldsymbol{A}_S^{(t+1)}, \boldsymbol{B}_S\Big). \tag{4}$$

Accordingly, the factor-conditioned second-order curvatures are

$$\boldsymbol{K}_A^{(t)} \propto \Big(\boldsymbol{B}_S^{(t)} \boldsymbol{H} \boldsymbol{B}_S^{(t)\top}\Big) \otimes \boldsymbol{G}, \qquad \boldsymbol{K}_B^{(t+1)} \propto \boldsymbol{H} \otimes \Big(\boldsymbol{A}_S^{(t+1)\top} \boldsymbol{G} \boldsymbol{A}_S^{(t+1)}\Big). \tag{5}$$

Hence, the sparsification of each factor incorporates both the inherited two-sided weighting and the latest realization of the opposite factor. This distinguishes the proposed calibration from independently sparsifying the two dense factors. For a parameter w_q and its corresponding factor-conditioned curvature $\boldsymbol{K}$, the standard OBS approximation gives

$$\Delta \boldsymbol{w} = -\frac{w_q}{[\boldsymbol{K}^{-1}]_{qq}} \boldsymbol{K}^{-1} \boldsymbol{e}^{(q)}, \qquad \Delta \mathcal{L}_q \approx \frac{w_q^2}{2[\boldsymbol{K}^{-1}]_{qq}}, \tag{6}$$

where the predicted loss increase provides the second-order importance criterion for $N{:}M$ mask selection. Detailed derivations are provided in Appendix A.2.

To reduce calibration overhead, NMP-SVD omits the sequential intra-block compensation commonly used in OBS-based pruning. Instead, each processing block is sparsified directly under the factor-conditioned criterion, and the accumulated approximation error is propagated to subsequent blocks for inter-block compensation. The alternating updates are repeated for multiple iterations, allowing later updates of $\boldsymbol{A}_S$ and $\boldsymbol{B}_S$ to further correct earlier sparsification errors. Residual errors are subsequently handled by the block-wise backfitting strategy introduced in Section 4.3.

4.3 Sparse Low-Rank Construction and Error Correction

Dense Initialization. NMP-SVD initializes the sparse solver from the dense low-rank solution of the underlying compressor. If native factors are unavailable, the weighted rank-r factors are recovered from $\overline{\boldsymbol{W}} \approx \boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}_r^\top$ as $\boldsymbol{A}_D = \boldsymbol{G}^{-1/2} \boldsymbol{U}_r \boldsymbol{\Sigma}_r^{1/2}$ and $\boldsymbol{B}_D = \boldsymbol{\Sigma}_r^{1/2} \boldsymbol{V}_r^\top \boldsymbol{H}^{-1/2}$. The dense solution initializes dual-factor calibration and is retained as the candidate for adaptive selection.

Residual-Guided Expansion. After obtaining the calibrated sparse base solution $\boldsymbol{S}_r = \boldsymbol{A}_S \boldsymbol{B}_S$, MP-SVD expands the factor width from its actual residual, $\boldsymbol{R}^{(0)} = \boldsymbol{W} - \boldsymbol{S}_r = (\boldsymbol{W} - \boldsymbol{D}_r) + (\boldsymbol{D}_r - \boldsymbol{S}_r)$, which jointly contains low-rank truncation and sparsification errors. For the j-th width-b_j expansion block, the current residual is decomposed in weighted coordinates as $\overline{\boldsymbol{R}}^{(j-1)} = \boldsymbol{G}^{1/2} \boldsymbol{R}^{(j-1)} \boldsymbol{H}^{1/2} \approx \boldsymbol{U}_j \boldsymbol{\Sigma}_j \boldsymbol{V}_j^\top$. The resulting block is mapped back to the original coordinates, calibrated using the same dual-factor $N{:}M$-OBS solver, and used to update the residual as $\boldsymbol{R}^{(j)} = \boldsymbol{R}^{(j-1)} - \boldsymbol{A}_S^{(j)} \boldsymbol{B}_S^{(j)}$. After J expansions, the factor width becomes $r' = r + \sum_{j=1}^{J} b_j$. Unlike extending the spectral tail of a dense decomposition, this strategy allocates new projections directly to information not captured by the current sparse realization.

Block-Wise Backfitting. Because residual blocks are constructed sequentially, early blocks and dense initialization block are not optimized with respect to the final expanded representation. NMP-SVD therefore applies lightweight block-wise backfitting. Defining the base sparse solution as block 0, $(\boldsymbol{A}_S^{(0)}, \boldsymbol{B}_S^{(0)}) = (\boldsymbol{A}_S, \boldsymbol{B}_S)$, block j is recalibrated against the conditional residual

$$\boldsymbol{R}_{-j} = \boldsymbol{W} - \sum_{\substack{i=0 \\ i \neq j}}^{J} \boldsymbol{A}_S^{(i)} \boldsymbol{B}_S^{(i)}. \tag{7}$$

The update is accepted only if it reduces the weighted reconstruction loss, correcting sequential-expansion interactions.

4.4 Plug-and-Play Integration and Adaptive Realization

NMP-SVD operates after weighting and rank allocation of the base SVD compressor. For each linear layer l, it retains the rank-r_l dense solution $\boldsymbol{D}_l$ before residual-guided expansion and constructs the expanded $N{:}M$ sparse realization $\boldsymbol{S}_l$ from the same initialization. Thus, adaptive selection requires no extra dense decomposition. Motivated by layer-wise variation in projection tolerance, NMP-SVD selects between $\boldsymbol{D}_l$ and $\boldsymbol{S}_l$ using local reconstruction errors E_D and E_S on the calibration set:

$$\widehat{\boldsymbol{W}}_l = \begin{cases} \boldsymbol{S}_l, & E_S < E_D, \\ \boldsymbol{D}_l, & \text{otherwise.} \end{cases} \tag{8}$$

Thus, sparsity-sensitive layers retain the original dense SVD solution, while the remaining layers adopt sparse SVD. For $W \in \mathbb{R}^{m\times n}$, a dense rank-$r$ factorization stores $r(m+n)$ factor values. An $N{:}M$ sparse width-r' realization retains $\frac{N}{M}r'(m+n)$ nonzero values, but additionally requires metadata to encode the structured sparse pattern. To account for this overhead consistently, we report an *equivalent parameter count* (P_{eq}), obtained by converting the metadata storage into an equivalent number of FP16 parameters and adding it to the stored factor values. We define $\rho = r'/r$ as the projection-width expansion ratio controlling the quality–efficiency trade-off.

5 Experiments

5.1 Experimental Setup

Models and Baselines. We evaluate NMP-SVD on LLaMA2-13B/70B, LLaMA3-8B, and Qwen3-8B/32B under two settings. First, to isolate sparse projection realization, we use uniform base-rank allocation with ASVD (Yuan et al., 2023) without STRS, SVD-LLM (Wang et al., 2025c), and KFAC-SVD (Thoma et al., 2026), with adaptive dense–sparse selection disabled. Second, we test compatibility with advanced non-uniform compressors, including SVD-LLM V2 (Wang et al., 2025b), KFAC-LEMS (Thoma et al., 2026), and Zero-Sum SVD (Abbasi et al., 2026), while preserving their original weighting and rank-allocation procedures.

Evaluation Setup. We evaluate perplexity on WikiText2 (Merity et al., 2017) and C4 (Raffel et al., 2020). Zero-shot performance is measured with LM-Evaluation-Harness (Gao et al., 2021) on ARC-c, ARC-e (Clark et al., 2018), WinoGrande (WG) (Sakaguchi et al., 2020), BoolQ (BQ) (Clark et al., 2019), HellaSwag (HS) (Zellers et al., 2019), OpenBookQA (OBQA) (Mihaylov et al., 2018), PIQA (Bisk et al., 2020) and MMLU (Hendrycks et al., 2020).

Compression and Implementation. We evaluate base parameter-reduction ratios of 20%, 40%, and 60% in the controlled study, and 20% and 40% in the non-uniform setting. Here, the ratio refers to the reduction used to determine the base Dense-SVD rank before NMP-SVD realization; NMP-SVD preserves this rank allocation and subsequently expands the sparse representation by ρ. Throughout the paper, we report model size using the equivalent parameter count (P_{eq}), and NMP-SVD uses $2{:}4$ structured sparsity by default. Unless noted otherwise, all methods use the same 128 WikiText2 calibration sequences of length 2048. Experiments use PyTorch and Hugging Face Transformers on one NVIDIA A800 GPU, except LLaMA2-70B, which uses two GPUs.

5.2 General Effectiveness under Uniform Rank Allocation

Table 1 isolates the effect of NMP-SVD under fixed uniform base-rank allocation. Across ASVD, SVD-LLM, and KFAC-SVD, NMP-SVD generally achieves lower WikiText2 perplexity and higher zero-shot accuracy across model families and compression levels, with the gains becoming more pronounced under stronger compression. Importantly, these improvements are achieved with a lower equivalent parameter count than the corresponding dense low-rank baselines, demonstrating that the gains do not arise from a larger storage budget. The two expansion ratios represent different quality–efficiency operating points: $\rho = 1.5$ generally provides stronger recovery, whereas $\rho = 1.4$ yields a more compact realization. Overall, the results show that NMP-SVD improves sparse low-rank representation across activation-, whitening-, and Fisher-weighted SVD backbones without relying

Table 1: WikiText2 (WT2)↓, average zero-shot accuracy (%)↑, and equivalent parameter count (P_{eq})↓ across models and SVD backbones under uniform base-rank allocation. NMP-SVD uses 2:4 sparse realizations without adaptive dense–sparse selection. Best results are bolded.

Ratio	Method	ρ	Llama2-13B			Llama3-8B			Qwen3-8B			Qwen3-32B		
			P_{eq}↓	WT2↓	AVG↑	P_{eq}↓	WT2↓	AVG↑	P_{eq}↓	WT2↓	AVG↑	P_{eq}↓	WT2↓	AVG↑
0%	Dense	–	13B	4.88	67.19	8B	6.13	69.34	8.2B	9.71	69.86	32.8B	7.61	74.51
20%	ASVD	–	10.5B	10.04	58.19	6.6B	521.3	37.98	6.8B	30.83	52.53	26.5B	18.88	63.48
	+NMP-SVD	1.4	**8.3B**	6.64	63.01	**5.4B**	16.85	58.66	**5.6B**	15.18	61.20	**21.2B**	11.26	68.42
	+NMP-SVD	1.5	8.9B	**6.08**	**63.83**	5.8B	**10.98**	**60.51**	5.9B	**13.33**	**64.30**	22.6B	**9.22**	**71.46**
	SVD-LLM	–	10.5B	6.92	57.77	6.6B	15.60	45.49	6.8B	15.07	57.67	26.5B	11.09	66.95
	+NMP-SVD	1.4	**8.3B**	6.41	58.34	**5.4B**	12.73	51.04	**5.6B**	13.36	61.82	**21.2B**	9.99	69.05
	+NMP-SVD	1.5	8.9B	**6.03**	**60.24**	5.8B	**10.48**	**54.91**	5.9B	**12.63**	**62.37**	22.6B	**9.35**	**69.89**
	KFAC-SVD	–	10.5B	6.67	57.66	6.6B	13.67	48.25	6.8B	13.50	58.41	26.5B	9.92	67.07
	+NMP-SVD	1.4	**8.3B**	6.38	58.71	**5.4B**	12.49	53.71	**5.6B**	12.98	61.47	**21.2B**	9.34	69.41
	+NMP-SVD	1.5	8.9B	**5.98**	**60.58**	5.8B	**10.40**	**55.84**	5.9B	**11.90**	**63.13**	22.6B	**8.86**	**70.15**
40%	ASVD	–	7.9B	55.82	46.26	5.2B	1.0e5	33.88	5.4B	251.5	40.93	20.3B	68.65	51.79
	+NMP-SVD	1.4	**6.3B**	14.61	52.07	**4.3B**	1958	37.17	**4.5B**	48.92	47.04	**16.3B**	25.63	57.68
	+NMP-SVD	1.5	6.7B	**11.05**	**55.41**	4.6B	**683.0**	**41.15**	4.8B	**32.78**	**51.59**	17.4B	**19.55**	**61.63**
	SVD-LLM	–	7.9B	11.79	45.24	5.2B	87.76	34.40	5.4B	28.00	38.37	20.3B	17.50	53.67
	+NMP-SVD	1.4	**6.3B**	10.06	47.23	**4.3B**	56.42	35.11	**4.5B**	21.94	48.72	**16.3B**	14.51	59.49
	+NMP-SVD	1.5	6.7B	**8.79**	**49.96**	4.6B	**32.10**	**40.11**	4.8B	**19.22**	**51.89**	17.4B	**12.69**	**62.14**
	KFAC-SVD	–	7.9B	11.58	46.10	5.2B	65.00	35.28	5.4B	23.23	42.02	20.3B	14.65	55.53
	+NMP-SVD	1.4	**6.3B**	10.06	48.26	**4.3B**	50.67	37.10	**4.5B**	19.71	47.76	**16.3B**	12.85	59.19
	+NMP-SVD	1.5	6.7B	**8.73**	**50.53**	4.6B	**28.96**	**42.95**	4.8B	**17.31**	**52.00**	17.4B	**11.70**	**62.07**
60%	ASVD	–	5.4B	1.7e4	35.23	3.8B	7.5e5	**37.11**	4B	2.8e4	35.19	14B	1311	37.92
	+NMP-SVD	1.4	**4.3B**	1992	36.47	**3.3B**	2.9e5	35.91	**3.4B**	2269	36.55	**11.4B**	253.3	40.73
	+NMP-SVD	1.5	4.6B	**596.8**	**37.30**	3.4B	**4.9e4**	34.48	3.6B	**815.9**	**37.08**	12.1B	**142.9**	**42.61**
	SVD-LLM	–	5.4B	44.93	34.88	3.8B	819.8	32.87	4B	182.0	33.45	14B	72.77	34.86
	+NMP-SVD	1.4	**4.3B**	36.52	35.27	**3.3B**	901.7	**33.55**	**3.4B**	142.1	33.77	**11.4B**	45.55	36.37
	+NMP-SVD	1.5	4.6B	**28.55**	**35.99**	3.4B	**748.5**	33.27	3.6B	**115.7**	**34.35**	12.1B	**33.58**	**38.50**
	KFAC-SVD	–	5.4B	43.47	34.72	3.8B	651.4	33.38	4B	126.1	33.85	14B	46.85	36.05
	+NMP-SVD	1.4	**4.3B**	34.80	35.35	**3.3B**	598.7	**33.71**	**3.4B**	103.1	34.37	**11.4B**	30.26	38.93
	+NMP-SVD	1.5	4.6B	**28.13**	**35.62**	3.4B	**492.2**	33.11	3.6B	**97.73**	**35.37**	12.1B	**24.49**	**43.24**

on heterogeneous rank allocation or adaptive layer selection. Detailed results for uniform rank allocation, including LLaMA2-70B with adaptive selection, are provided in Appendices F.1 and F.2.

5.3 Compatibility with Non-Uniform Rank Allocation

We next apply NMP-SVD after the original layer-wise rank allocation of SVD-LLM V2, KFAC-LEMS, and Zero-Sum SVD; for Zero-Sum SVD, we use the $5\times$ truncate–correct–re-truncate setting. As shown in Table 2, NMP-SVD remains effective with heterogeneous rank allocation, generally reducing the equivalent parameter count and improving prefill end-to-end speedup while maintaining or improving C4 perplexity and zero-shot accuracy. Speedup is not strictly monotonic with ρ: although $\rho = 1.5$ increases the cost of each sparse candidate, its improved reconstruction allows more layers to retain the sparse realization, partially offsetting the higher per-layer cost. Zero-Sum SVD is an exception on WT2 in several settings, where its original dense solution remains better; this may reflect stronger adaptation of its repeated correction procedure to the WikiText2-based calibration objective. Nevertheless, NMP-SVD performs better on C4 and zero-shot evaluation in most corresponding settings while providing stronger compression and prefill efficiency. Detailed results of non-uniform rank allocation are provided in Appendix F.3.

5.4 Ablation Study

Method components and adaptive realization. Table 3 isolates the impact of NMP-SVD components, with all variants evaluated under the same all-sparse realization to avoid interference from adaptive selection (baseline is KFAC-SVD with $\rho = 1.4$). We compare magnitude-based $N{:}M$ masking in place of OBS calibration (*Magnitude-*$N{:}M$), direct SparseGPT sparsification of dense

Table 2: WikiText2/C4 perplexity↓, average zero-shot accuracy (%)↑, end-to-end prefill speedup↑, and equivalent parameter count (P_{eq}↓) for non-uniform SVD compressors with and without NMP-SVD. Speedup is measured at batch size 4 and sequence length 2048 on RTX 3090. Best results within each baseline group are bolded.

Ratio	Method	ρ	Llama3-8B					Qwen3-8B				
			Speedup	P_{eq}↓	WT2↓	C4↓	AVG↑	Speedup	P_{eq}↓	WT2↓	C4↓	AVG↑
0%	Dense	–	1×	8B	6.13	9.26	69.34	1×	8.2B	9.71	15.23	69.86
20%	Zero-Sum SVD	–	1.11×	6.6B	12.51	33.58	51.76	1.10×	6.8B	**11.45**	23.41	65.91
	+NMP-SVD	1.4	**1.24**×	**6.2B**	11.78	23.96	54.35	**1.24**×	**6.2B**	11.69	20.48	66.06
	+NMP-SVD	1.5	1.20×	6.3B	**10.78**	**20.55**	**56.72**	1.23×	6.4B	11.53	**19.21**	**66.70**
	SVD-LLM-v2	–	1.00×	6.6B	15.98	90.93	44.34	0.95×	6.8B	14.94	37.22	57.06
	+NMP-SVD	1.4	**1.32**×	**5.6B**	12.52	37.09	52.62	**1.29**×	**5.9B**	13.40	29.63	61.27
	+NMP-SVD	1.5	1.25×	5.8B	**10.43**	**26.23**	**57.35**	1.23×	6B	**12.41**	**22.28**	**62.61**
	KFAC-LEMS	–	1.13×	6.6B	9.33	17.46	60.85	1.13×	6.8B	11.30	19.71	65.82
	+NMP-SVD	1.4	**1.20**×	**6.3B**	8.84	16.02	62.27	**1.21**×	**6.4B**	10.99	18.55	66.32
	+NMP-SVD	1.5	1.19×	6.4B	**8.45**	**14.98**	**63.11**	1.20×	6.5B	**10.72**	**17.82**	**67.19**
40%	Zero-Sum SVD	–	1.28×	5.2B	**52.16**	455.6	33.55	1.26×	5.4B	23.21	95.50	47.50
	+NMP-SVD	1.4	**1.62**×	**4.5B**	116.3	564.3	34.01	**1.61**×	**4.6B**	22.52	84.69	47.39
	+NMP-SVD	1.5	1.54×	4.7B	79.87	**445.2**	**34.62**	1.53×	4.8B	**19.99**	**68.02**	**51.15**
	SVD-LLM-v2	–	1.26×	5.2B	114.8	1117	34.04	1.25×	5.4B	32.31	140.6	37.53
	+NMP-SVD	1.4	**1.64**×	**4.4B**	79.36	510.6	34.01	**1.58**×	**4.6B**	22.50	86.84	48.08
	+NMP-SVD	1.5	1.58×	4.6B	**46.19**	**287.3**	**36.41**	1.53×	4.8B	**19.81**	**68.11**	**51.00**
	KFAC-LEMS	–	1.34×	5.2B	25.49	97.69	43.08	1.37×	5.4B	20.59	62.35	47.77
	+NMP-SVD	1.4	1.44×	**4.8B**	21.72	69.31	46.45	1.50×	**4.9B**	20.89	55.10	50.88
	+NMP-SVD	1.5	**1.51**×	4.9B	**18.44**	**52.44**	**48.31**	**1.50**×	5.1B	**18.32**	**47.47**	**52.62**

Table 3: Component ablation on LLaMA3-8B.

Variant	20%		40%	
	WT2 ↓	AVG ↑	WT2 ↓	AVG ↑
Magnitude-N:M	18.23	46.58	104.8	34.76
Direct SparseGPT	29.07	43.31	145.5	34.46
w/o Backfit	15.41	49.49	90.98	35.29
w/o Residual	12.64	52.72	53.04	36.16
NMP-SVD	**12.49**	**53.71**	**50.67**	**37.10**

Table 4: Effect of adaptive dense–sparse selection.

Model	Ratio	All Sparse		Adaptive	
		WT2 ↓	AVG ↑	WT2 ↓	AVG ↑
LLaMA2-13B	20%	6.38	58.71	**6.28**	**59.26**
	40%	10.06	48.26	**9.88**	**49.30**
LLaMA3-8B	20%	12.49	53.71	**12.18**	**54.20**
	40%	**50.67**	**37.10**	51.21	36.54

width-ρr low-rank factors (*Direct SparseGPT*), removal of block-wise backfitting (*w/o Backfit*), and expansion from the dense residual rather than the sparse residual (*w/o Residual*). All variants yield higher WT2 perplexity and lower zero-shot accuracy, particularly under stronger compression, supporting the benefits of factor-aware sparse calibration, sparse-residual expansion, and backfitting. Table 4 separately evaluates the adaptive dense–sparse selector. It improves model quality in most settings over the all-sparse realization. However, since the selector is based on local calibration reconstruction error, it is not guaranteed to improve downstream metrics in every case. Table 4 reflects this limitation. Detailed results of adaptive dense–sparse SVD selection are provided in Appendix C.

Iteration efficiency and quantization compatibility. Figure 4 evaluates dual-factor calibration on representative LLaMA3-8B linear layers across multiple depths and projection types at 40% compression ratio ($\rho = 1.4$, baseline is KFAC-SVD), with reconstruction error normalized per layer before aggregation. Additional iterations reduce reconstruction error, but the marginal gain diminishes while execution time grows approximately linearly. We therefore use three iterations by default and two for Qwen3-32B and LLaMA2-70B to limit offline compression cost. Table 5 further shows that NMP-SVD is compatible with 8-bit RTN quantization. INT8 approximately halves the factor-value storage relative to FP16 NMP-SVD while retaining a clear advantage over the KFAC-SVD baseline. INT8 leaves perplexity nearly unchanged and preserves zero-shot accuracy, indicating that NMP-SVD remains compatible with post-compression quantization.

Additional results on calibration size and dataset choice are provided in Appendix B, while the impact of different N:M sparsity patterns is reported in Appendix B.3.

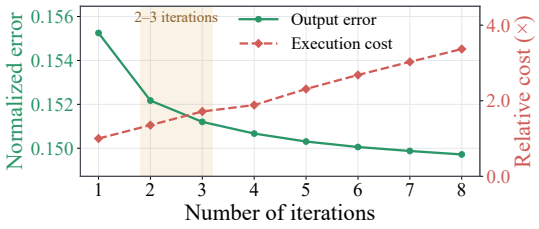
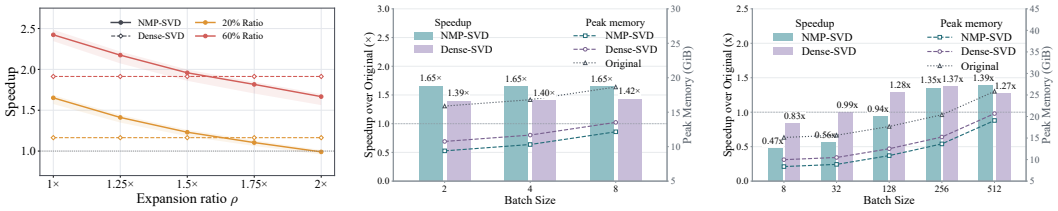


Figure 4: Effect of dual-factor calibration iterations.

Table 5: Compatibility with 8-bit RTN quantization on LLaMA2-13B with adaptive dense-sparse selection.

Method	20%		40%	
	WT2↓	AVG↑	WT2↓	AVG↑
KFAC-SVD	6.67	57.66	11.58	46.10
+NMP-SVD	**6.28**	59.26	**9.88**	49.30
+NMP-SVD + INT8	6.28	**59.43**	9.89	**49.34**



(a) Single-layer speedup. (b) Prefill speedup and memory. (c) Decode speedup and memory.

Figure 5: Inference efficiency of NMP-SVD on an NVIDIA RTX 3090. (a) Single-layer speedup under different expansion ratios. (b–c) End-to-end speedup and peak memory of LLaMA3-8B during prefill (sequence length 2048) and decoding (context length 128), respectively.

5.5 Efficiency Analysis

Figure 5a evaluates the single-layer efficiency of 2:4 sparse low-rank projections relative to Dense SVD on LLaMA2-13B at 20% and 60% compression across different expansion ratios ρ. Increasing ρ provides greater projection capacity but also increases sparse GEMM cost. At $\rho = 1.5$, NMP-SVD retains approximately $0.5 \times 1.5 = 0.75$ of the nonzero parameters of Dense-r while achieving comparable or slightly higher speed, providing a favorable quality–efficiency operating point.

Figures 5b and c report end-to-end results on LLaMA3-8B. During prefill, NMP-SVD consistently achieves higher speedup and lower peak memory than Dense-SVD across the tested batch sizes. During autoregressive decoding, its efficiency is more batch-dependent: NMP-SVD is slower at small batches, approaches Dense-SVD as batch-level parallelism increases, and surpasses it at the largest tested batch size. This crossover indicates that sparse-kernel overhead dominates at low concurrency but is increasingly amortized at larger batches. NMP-SVD also maintains a clear memory advantage across the evaluated decoding settings. Overall, NMP-SVD provides robust prefill acceleration and becomes competitive with Dense-SVD at high concurrency and surpasses it at the largest tested batch size, while reducing memory usage in both inference stages. Details of the inference evaluation setup are provided in Appendix E.

Compression overhead results for NMP-SVD are provided in Appendix D.

6 Conclusion

In this work, we revisit the assumption that low-rank projections in SVD-based LLM compression must remain fully dense. Controlled perturbation experiments show that low-rank representations can tolerate substantial directional approximation, with tolerance varying across models and linear layers. Motivated by this observation, we propose NMP-SVD, a plug-and-play projection-realization framework that combines dual-factor N:M-OBS calibration, sparse-residual projection expansion, and adaptive dense–sparse selection while preserving the weighting and rank-allocation strategies of existing SVD compressors. Experiments across multiple LLM families show improved quality–compression trade-offs, more compact low-rank representations, higher prefill efficiency, and lower memory usage during both prefill and decoding. These results establish projection realization as a complementary design dimension for SVD-based LLM compression.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing, manuscript polishing, and code development and debugging. We have not used generative AI tools to autonomously conduct experiments, generate experimental results, or make final scientific judgments, and other disclosure categories are not applicable to this work. Additionally, we used generative AI tools to improve the clarity and organization of selected passages in the manuscript. We have reviewed all AI-assisted work. All AI-assisted text was manually checked and revised by the authors, and AI-assisted code was inspected, executed, and validated for correctness before being used in the experiments. All experimental results, analyses, and conclusions were independently verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Ali Abbasi, Chayne Thrash, Haoran Qin, Shansita Sharma, Sepehr Seifi, and Soheil Kolouri. Zero sum SVD: Balancing loss sensitivity for low rank LLM compression. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=tAdsDOkRBw.

Guangji Bai, Yijiang Li, Chen Ling, Kibaek Kim, and Liang Zhao. SparseLLM: Towards global pruning of pre-trained language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=oXHyYHp4Zb.

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.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2924–2936, 2019.

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.

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.

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

Leo Gao, Jonathan Tow, Stella Biderman, Shawn Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jasmine Hsu, Kyle McDonell, Niklas Muennighoff, et al. A framework for few-shot language model evaluation. *Version v0. 0.1. Sept*, 10:8–9, 2021.

Tianteng Gu, Bei Liu, Bo Xiao, Ke Zeng, Jiacheng Liu, and Yanmin Qian. Denoiserotator: Enhance pruning robustness for llms via importance concentration. *Advances in Neural Information Processing Systems*, 38:64037–64067, 2026.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Mihee Lee, Haipeng Shen, Jianhua Z Huang, and James S Marron. Biclustering via sparse singular value decomposition. *Biometrics*, 66(4):1087–1095, 2010.

Hongyi Liu, Rajarshi Saha, Zhen Jia, Youngsuk Park, Jiaji Huang, Shoham Sabach, Yu-Xiang Wang, and George Karypis. PROXSPARSE: REGULARIZED LEARNING OF SEMI-STRUCTURED SPARSITY MASKS FOR PRETRAINED LLMS. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=zkxe5vASi8.

Mehdi Makni, Xiang Meng, and Rahul Mazumder. 3basil: An algorithmic framework for sparse plus low-rank compression of llms. *Advances in Neural Information Processing Systems*, 38: 173917–173947, 2026.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

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 Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2381–2391, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1260. URL https://aclanthology.org/D18-1260/.

Mohammad Mozaffari, Amir Yazdanbakhsh, and Maryam Mehri Dehnavi. Slim: One-shot quantization and sparsity with low-rank approximation for llm weight compression. *arXiv preprint arXiv:2410.09615*, 2024.

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.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8732–8740, 2020.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PxoFut3dWW.

Moritz Thoma, Maximilian Groezinger, Maximilian Forstenhäusler, Emad Aghajanzadeh, Manoj Rohit Vemparala, Christos Anagnostopoulos, Pierpaolo Mori, Nael Fasfous, Alexander Frickenstein, Daniel Mueller-Gritschneder, and Ulf Schlichtmann. Advancing SVD-based LLM compression via layer-wise error model search. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=IjIgNPFuCt.

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*, 2023.

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

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

Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. Svd-llm: Truncation-aware singular value decomposition for large language model compression. In *International Conference on Learning Representations*, volume 2025, pp. 19299–19319, 2025c.

Canwen Xu and Julian McAuley. A survey on model compression and acceleration for pretrained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10566–10575, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zhihang Yuan, Yuzhang Shang, Yue Song, Dawei Yang, 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.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, 12: 1556–1577, 2024.

Lancheng Zou, Shuo Yin, Zehua Pei, Tsung-Yi Ho, Farzan Farnia, and Bei Yu. Permllm: Learnable channel permutation for n: M sparse large language models. *Advances in Neural Information Processing Systems*, 38:20183–20205, 2026.

A PROOFS

A.1 DERIVATION OF THE APPROXIMATE PROJECTION-TOLERANCE SCALE

We provide an idealized analysis for the projection-tolerance scale used in Section 3. Let $\boldsymbol{W} = \sum_{i=1}^{d} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ and define the rank-one components $\boldsymbol{P}_i = \boldsymbol{u}_i \boldsymbol{v}_i^\top$. For $k = \rho r$, let $S_k = \sum_{i=1}^{k} \sigma_i^2$. The reconstruction errors of the exact Dense-r and Dense-k approximations are $E_r = \sum_{i>r} \sigma_i^2$ and $E_k = \sum_{i>k} \sigma_i^2$, respectively, so rank expansion recovers

$$E_r - E_k = S_k - S_r. \tag{9}$$

To preserve the rank-one structure of each projection component, we perturb the singular-vector direction rather than the component directly. Specifically, we keep $\boldsymbol{u}_i$ fixed and construct $\widetilde{\boldsymbol{v}}_i = c\boldsymbol{v}_i + \sqrt{1-c^2}\,\boldsymbol{q}_i$, where $\|\boldsymbol{q}_i\|_2 = 1$, $\boldsymbol{q}_i$ lies in the orthogonal complement of the retained right-singular subspace, and the $\boldsymbol{q}_i$ are mutually orthogonal. The perturbed component is therefore $\widetilde{\boldsymbol{P}}_i = \boldsymbol{u}_i \widetilde{\boldsymbol{v}}_i^\top$, which remains rank one.

Because $\|\boldsymbol{P}_i\|_F = \|\widetilde{\boldsymbol{P}}_i\|_F = 1$,

$$\frac{\langle \boldsymbol{P}_i, \widetilde{\boldsymbol{P}}_i \rangle_F}{\|\boldsymbol{P}_i\|_F \|\widetilde{\boldsymbol{P}}_i\|_F} = \boldsymbol{v}_i^\top \widetilde{\boldsymbol{v}}_i = c, \tag{10}$$

so c is equivalently the Frobenius-space cosine similarity between the original and perturbed rank-one components. Moreover, $\|\widetilde{\boldsymbol{P}}_i - \boldsymbol{P}_i\|_F^2 = 2(1-c)$.

The perturbed rank-k representation $\widetilde{\boldsymbol{D}}_k = \sum_{i=1}^{k} \sigma_i \widetilde{\boldsymbol{P}}_i$ therefore remains within the rank-k factorization class. Under the orthogonality construction above, its reconstruction error is

$$\|\boldsymbol{W} - \widetilde{\boldsymbol{D}}_k\|_F^2 = E_k + 2(1-c)S_k. \tag{11}$$

Requiring the perturbed rank-k representation to remain no worse than Dense-r gives

$$E_k + 2(1-c)S_k \le E_r, \tag{12}$$

and hence

$$\tau_{\mathrm{spec}}(r, \rho) = 1 - c \le \frac{1}{2}\left(1 - \frac{S_r}{S_{\rho r}}\right). \tag{13}$$

Under the idealized approximation $S_{\rho r} \approx \rho S_r$, this becomes

$$\tau_{\text{approx}}(\rho) \approx \frac{\rho - 1}{2\rho}. \tag{14}$$

This gives reference values of approximately 0.10, 0.17, 0.21, and 0.25 for $\rho = 1.25$, 1.5, 1.75, and 2, respectively. These values are not universal bounds; the actual tolerance depends on the singular-value spectrum and therefore varies across layers and projection types.

A.2 Derivation of Dual-Factor N:M-OBS Calibration

Starting from the weighted reconstruction objective

$$\mathcal{L}(\boldsymbol{A}, \boldsymbol{B}) = \left\| \boldsymbol{G}^{1/2}(\boldsymbol{W} - \boldsymbol{A}\boldsymbol{B})\boldsymbol{H}^{1/2} \right\|_F^2, \tag{15}$$

fixing either factor makes the objective quadratic with respect to the other. Using $\mathrm{vec}(\boldsymbol{P}\boldsymbol{X}\boldsymbol{Q}) = (\boldsymbol{Q}^\top \otimes \boldsymbol{P})\,\mathrm{vec}(\boldsymbol{X})$, the Hessian with respect to $\mathrm{vec}(\boldsymbol{A})$ is, up to a constant,

$$\boldsymbol{K}_A \propto (\boldsymbol{B}\boldsymbol{H}\boldsymbol{B}^\top) \otimes \boldsymbol{G}, \tag{16}$$

while fixing $\boldsymbol{A}$ gives

$$\boldsymbol{K}_B \propto \boldsymbol{H} \otimes (\boldsymbol{A}^\top \boldsymbol{G}\boldsymbol{A}). \tag{17}$$

Thus, the sparsification metric of each factor depends on both the inherited two-sided weighting and the current realization of the opposite factor.

For the factor currently being optimized, let $\boldsymbol{w}$ denote its vectorized parameters and $\boldsymbol{K}$ the corresponding factor-conditioned curvature. The local second-order loss variation is $\Delta\mathcal{L} \approx \frac{1}{2}\Delta\boldsymbol{w}^\top \boldsymbol{K}\Delta\boldsymbol{w}$. If parameter w_q is constrained to zero, $\boldsymbol{e}^{(q)\top}\Delta\boldsymbol{w} = -w_q$. Solving this constrained quadratic problem gives the standard OBS update

$$\Delta\boldsymbol{w} = -\frac{w_q}{[\boldsymbol{K}^{-1}]_{qq}}\boldsymbol{K}^{-1}\boldsymbol{e}^{(q)}, \qquad \Delta\mathcal{L}_q = \frac{w_q^2}{2[\boldsymbol{K}^{-1}]_{qq}}. \tag{18}$$

The predicted loss increase therefore provides the second-order importance criterion used to select the $M - N$ entries removed within each N:M group.

Because $\boldsymbol{K}_A$ depends on $\boldsymbol{B}$ and $\boldsymbol{K}_B$ depends on $\boldsymbol{A}$, NMP-SVD uses the Gauss–Seidel-style updates

$$\begin{aligned} \boldsymbol{K}_A^{(t)} &\propto (\boldsymbol{B}_S^{(t)}\boldsymbol{H}\boldsymbol{B}_S^{(t)\top}) \otimes \boldsymbol{G}, & \boldsymbol{A}_S^{(t+1)} &= \mathrm{OBS}_{N:M}(\boldsymbol{A}_S^{(t)}; \boldsymbol{K}_A^{(t)}), \\ \boldsymbol{K}_B^{(t+1)} &\propto \boldsymbol{H} \otimes (\boldsymbol{A}_S^{(t+1)\top}\boldsymbol{G}\boldsymbol{A}_S^{(t+1)}), & \boldsymbol{B}_S^{(t+1)} &= \mathrm{OBS}_{N:M}(\boldsymbol{B}_S^{(t)}; \boldsymbol{K}_B^{(t+1)}). \end{aligned} \tag{19}$$

This directly yields the factor-conditioned alternating calibration used in Section 4.2. In the practical implementation, Eq. 18 provides the local OBS criterion for mask selection, while NMP-SVD omits sequential intra-block compensation and propagates the accumulated approximation error across processing blocks. Subsequent alternating iterations and block-wise backfitting provide further error correction.

B Impact of Calibration Size and Dataset

B.1 Impact of Calibration Size

We study the sensitivity of NMP-SVD to calibration-set size on Qwen3-8B at 40% base parameter reduction with $\rho = 1.5$ under uniform rank allocation. We vary the number of WikiText2 calibration sequences from 32 to 512 while keeping all other compression settings fixed. As shown in Figure 6, increasing the calibration size consistently reduces perplexity on both WikiText2 and C4, with the largest gains occurring between 32 and 128 samples. Zero-shot accuracy follows a similar overall trend, improving substantially up to 128 samples and remaining relatively stable thereafter despite minor fluctuations. These results indicate diminishing returns beyond 128 calibration samples, supporting our default choice of 128 as a good balance between calibration quality and compression cost.

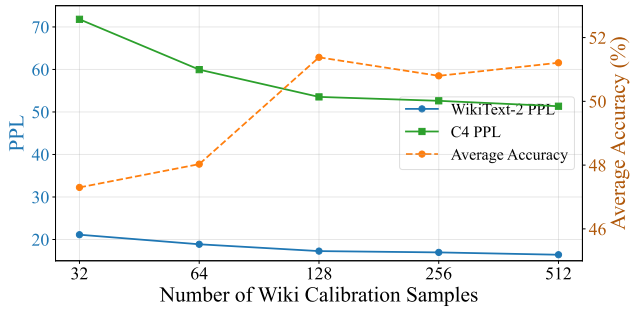


Figure 6: Effect of calibration-set size on Qwen3-8B at 40% ratio and $\rho = 1.5$.

Table 6: Effect of calibration dataset on Qwen3-8B at 40% compression ratio and $\rho = 1.5$. Calibration uses C4. Perplexity is reported on WikiText2 (WT2) and C4, while zero-shot task results and AVG are accuracies (%). The best result among the compressed configurations for each metric is highlighted in bold.

Ratio	**Method**	ρ	**Calib. N**	WT2↓	C4↓	ARC-c↑	ARC-e↑	WG↑	BQ↑	HS↑	OBQA↑	PIQA↑	MMLU↑	**AVG↑**
0%	Dense	–	–	9.71	15.23	56.48	80.93	67.80	86.57	74.93	41.40	77.80	72.97	69.86
40%	KFAC-SVD	–	128	72.05	44.39	27.56	39.73	60.77	65.14	45.87	30.80	66.87	25.49	45.28
	+NMP-SVD	1.5	128	**34.02**	**31.02**	**36.60**	**57.74**	**63.38**	**74.01**	**55.68**	**35.60**	**71.93**	**41.29**	**54.53**
	KFAC-SVD	–	256	72.68	42.72	28.24	39.86	60.22	65.23	47.00	31.00	66.76	25.78	45.51
	+NMP-SVD	1.5	256	34.08	**28.41**	**36.69**	**54.67**	**64.40**	**73.21**	**55.82**	**35.80**	**71.44**	**41.41**	**54.18**

B.2 Impact of Calibration Data

To examine the effect of calibration data, we replace WikiText2 with C4 and evaluate 128 and 256 calibration sequences under uniform rank allocation on Qwen3-8B at 40% base parameter reduction. As shown in Table 6, NMP-SVD consistently outperforms the KFAC-SVD baseline for both calibration sizes, substantially reducing perplexity on WT2 and C4 while improving zero-shot accuracy. Increasing the calibration size from 128 to 256 produces only modest changes, indicating that the gains of NMP-SVD are robust to the calibration-set size.

B.3 Impact of Different N:M Sparsity Patterns

Table 7 further examines alternative N:M patterns beyond the default 2:4 setting. The 1:4 pattern reduces the retained density from 50% to 25%, imposing a substantially stronger sparsity constraint and resulting in marked performance degradation. In contrast, 4:8 preserves the same 50% density as 2:4 while allowing greater flexibility within each sparsity group, yielding lower perplexity and higher zero-shot accuracy. These results are consistent with the expected trade-off between structural flexibility and reconstruction quality.

Table 7: Ablation of different N:M sparsity patterns for Llama3-8B with KFAC-SVD under uniform rank allocation. Perplexity is reported on WikiText2 (WT2) and C4, while zero-shot task results and AVG are accuracies (%). The best result within each sparsity group for each metric is highlighted in bold.

Ratio	Method	ρ	N:M	WT2↓	C4↓	ARC-c↑	ARC-e↑	WG↑	BQ↑	HS↑	OBQA↑	PIQA↑	MMLU↑	**AVG↑**
20%	KFAC-SVD	1.0	Dense	13.67	61.78	33.53	54.08	64.01	62.20	49.09	33.60	65.56	23.96	48.25
	+NMP-SVD	1.5	2:4	12.49	34.54	38.82	64.31	66.46	62.48	56.00	37.00	69.48	35.12	53.71
	+NMP-SVD	1.5	1:4	583.9	3845	24.32	27.78	51.54	37.83	26.19	26.20	49.89	23.28	33.38
	+NMP-SVD	1.5	4:8	**9.11**	**17.58**	**43.60**	**66.54**	**71.11**	**65.72**	**66.85**	**39.20**	**73.07**	**41.01**	**58.39**
40%	KFAC-SVD	1.0	Dense	65.00	563.0	23.29	33.04	53.04	37.83	28.50	29.20	53.10	24.21	35.28
	+NMP-SVD	1.5	2:4	28.96	157.9	26.37	45.58	58.48	61.07	37.09	31.20	60.83	23.00	42.95
	+NMP-SVD	1.5	1:4	977.1	3276	26.37	28.20	49.17	37.83	26.43	25.40	50.11	**25.07**	33.57
	+NMP-SVD	1.5	4:8	**21.98**	**105.3**	**27.56**	**49.24**	**61.33**	**62.48**	**40.52**	**32.80**	**62.08**	23.34	**44.92**

C Detailed Results for Adaptive Dense–Sparse SVD Selection

We further visualize the layer-wise decisions of the adaptive dense–sparse selector across models, base SVD methods, and expansion ratios. Each cell corresponds to one linear layer, with Dense SVD retained when its local reconstruction error is lower and Sparse SVD selected otherwise.

Effect of the base SVD method and model. The selection patterns differ substantially across both compression methods and model families. As shown in Figure 7 and Figure 8, under the same ρ, KFAC-SVD consistently retains more Dense SVD layers than SVD-LLM-v2. For example, on LLaMA3-8B, KFAC-SVD retains 70 dense layers at both $\rho = 1.4$ and 1.5, whereas SVD-LLM-v2 retains 59 and 12, respectively. The difference is even larger on Qwen3-8B, where KFAC-SVD retains 103 dense layers while SVD-LLM-v2 retains only 38 and 3. This behavior is consistent with the stronger two-sided weighting used by KFAC-SVD, which provides a more accurate dense reference and therefore makes the expanded sparse realization harder to outperform. The selection maps also vary across models and layer types, further confirming that sparsity tolerance is highly heterogeneous rather than uniform across the network.

Effect of the expansion ratio ρ. The expansion ratio ρ strongly influences the dense–sparse decision. As shown in Figure 9 and Figure 10, with KFAC-SVD at $\rho = 1.3$, most layers remain dense: 202 of 224 layers for LLaMA3-8B and 226 of 252 for Qwen3-8B. Increasing ρ to 1.4–1.5 reverses this pattern, with the majority of layers selecting Sparse SVD as the additional projection capacity improves sparse reconstruction. The transition is not strictly monotonic, however, as the number of dense layers increases slightly again at $\rho = 1.6$. This indicates that ρ controls the overall feasibility of sparse realization, while the final choice remains determined by the layer-specific reconstruction trade-off. Together, these results support adaptive selection instead of enforcing either dense or sparse realization uniformly across all layers.

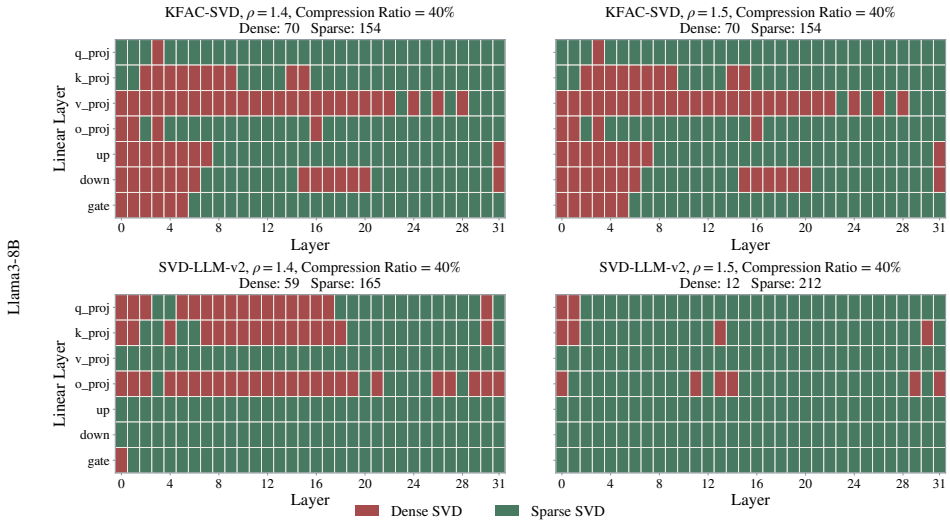


Figure 7: Layer-wise adaptive dense–sparse selection on LLaMA3-8B for KFAC-SVD and SVD-LLM-v2 at $\rho = 1.4$ and 1.5.

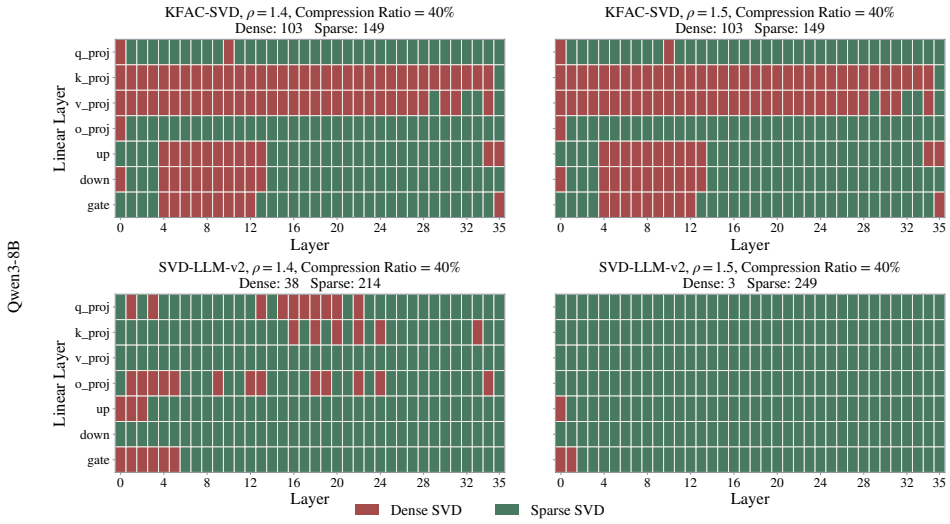


Figure 8: Layer-wise adaptive dense–sparse selection on Qwen3-8B for KFAC-SVD and SVD-LLM-v2 at $\rho = 1.4$ and 1.5.

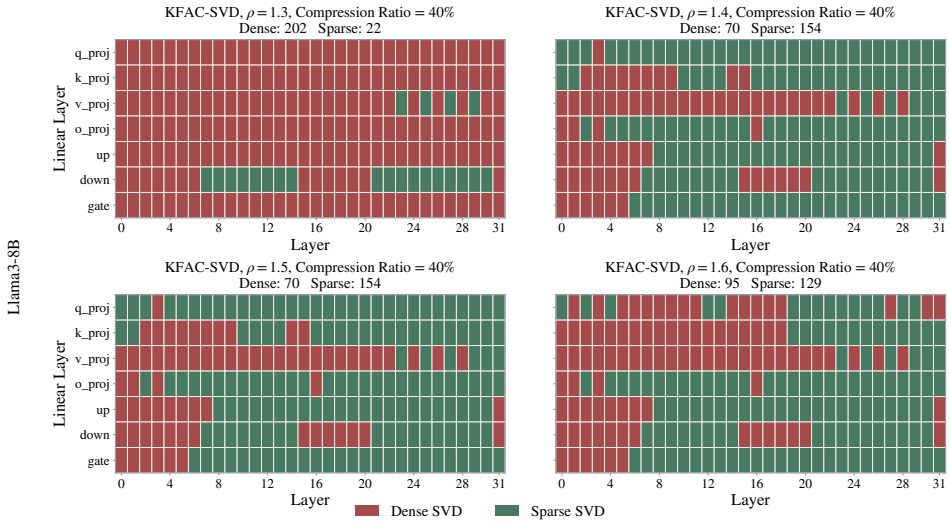


Figure 9: Effect of the expansion ratio ρ on layer-wise dense–sparse selection for KFAC-SVD on LLaMA3-8B.

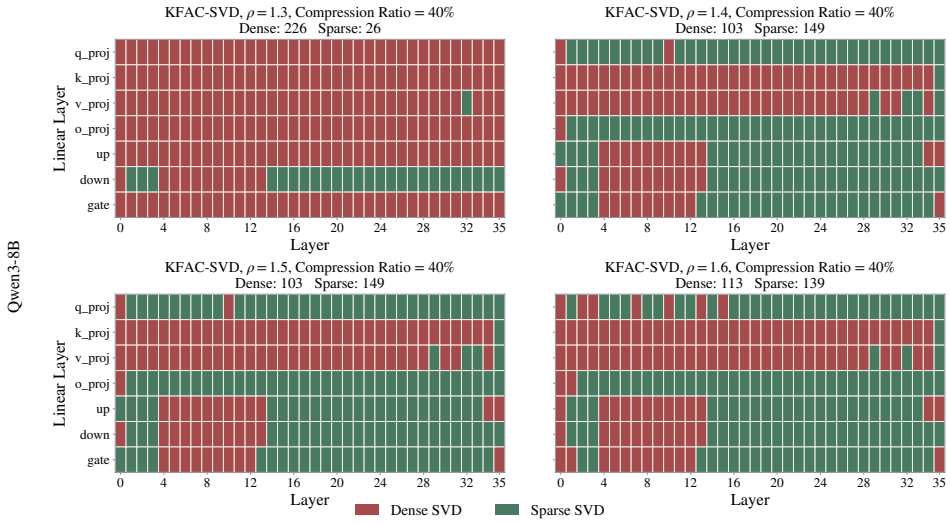


Figure 10: Effect of the expansion ratio ρ on layer-wise dense–sparse selection for KFAC-SVD on Qwen3-8B.

Table 8: Total compression time under uniform rank allocation at a 40% base parameter-reduction ratio. NMP-SVD uses 2:4 sparsity and $\rho = 1.5$. Times are reported in hours (h) and minutes (min).

Method	LLaMA2-13B	LLaMA3-8B	Qwen3-8B	Qwen3-32B	Llama2-70B
ASVD	20 min	9 min	8 min	29 min	1 h 42 min
+NMP-SVD	1 h 02 min	28 min	27 min	3 h 11 min	5 h 22 min
SVD-LLM	26 min	12 min	14 min	1 h 11 min	6 h 55 min
+NMP-SVD	1 h 08 min	35 min	33 min	4 h 01 min	11 h 55 min
KFAC-SVD	44 min	23 min	23 min	2 h 05 min	8 h 43 min
+NMP-SVD	2 h 04 min	1 h 07 min	1 h 01 min	8 h 11 min	16 h 01 min

D Compression Overhead

NMP-SVD introduces additional offline computation for factor-conditioned OBS calibration, alternating updates, residual expansion, and backfitting. Table 8 reports the total compression time under uniform rank allocation at a 40% base parameter-reduction ratio with $\rho = 1.5$. The relative overhead depends strongly on the underlying compressor because the reported time also includes backbone-specific weighting calibration. In particular, KFAC-SVD requires costly estimation of two-sided weighting matrices, whereas ASVD incurs relatively little overhead in collecting its weighting statistics. The ASVD results therefore more directly reflect the additional cost of NMP-SVD itself, which is approximately $3\times - 5\times$ that of the corresponding dense SVD pipeline for most model scales. The additional optimization is performed only once during offline compression and does not affect inference-time execution.

E Inference Measurement Protocol

Inference efficiency is measured on NVIDIA GeForce RTX 3090 GPUs with PyTorch 2.5.1 and CUDA 11.8. Dense and sparse computations use FP16; 2:4 sparse factors are implemented with PyTorch semi-structured tensors (`torch.sparse.to_sparse_semi_structured`), and attention uses the Flash SDP backend.

For prefill, we evaluate batch sizes of 2, 4, and 8 with a sequence length of 2048. For decoding, we use an initial context length of 128 and evaluate batch sizes of 8, 32, 128, 256, and 512, generating 32 tokens autoregressively with KV caching enabled. The shorter decoding context enables large-batch evaluation within GPU memory limits and allows us to examine throughput-oriented serving regimes. Inputs are generated directly on the GPU. Model construction, tokenization, and host-to-device transfer are excluded from timing.

Each configuration is warmed up for 10 runs and measured over 30 runs using CUDA events, with median latency reported. Speedup is measured relative to the original dense model. Peak memory is measured using `torch.cuda.max_memory_allocated()` and includes model parameters, activations, and the KV cache.

F Detail Experiments

F.1 Detail Results under Uniform Rank Allocation without Adaptive Dense-Sparse Selection

Detailed evaluation results for Llama2-13B, Llama3-8B, Qwen3-8B and Qwen3-32B are shown in Table 9, Table 10, Table 11 and Table 12, respectively. We report equivalent parameter count (P_{eq}), perplexity on WikiText2 (WT2) and C4, and zero-shot accuracy (%) on ARC-c, ARC-e, WinoGrande (WG), BoolQ (BQ), HellaSwag (HS), OpenBookQA (OBQA), PIQA, MMLU, and their average (AVG). The best result within each baseline group for each metric is highlighted in bold.

Table 9: Detailed evaluation results for Llama2-13B.

Ratio	Method	ρ	$P_{\text{eq}}\downarrow$	WT2↓	C4↓	ARC-c↑	ARC-e↑	WG↑	BQ↑	HS↑	OBQA↑	PIQA↑	MMLU↑	**AVG↑**
0%	Dense	–	13B	4.88	6.62	48.81	76.52	72.38	82.11	79.64	45.20	80.41	52.44	67.19
20%	ASVD	–	10.5B	10.04	13.11	38.65	64.39	68.59	75.60	65.93	40.20	74.16	37.97	58.19
	+NMP-SVD	1.4	**8.3B**	6.64	9.51	43.52	69.07	71.98	79.36	73.23	**43.80**	**77.80**	45.35	63.01
	+NMP-SVD	1.5	8.9B	**6.08**	**8.59**	**44.97**	**71.93**	**72.53**	**79.79**	**75.09**	42.60	77.75	**46.01**	**63.83**
	SVD-LLM	–	10.5B	6.92	14.53	38.99	67.34	**70.40**	75.44	62.93	40.80	72.63	33.64	57.77
	+NMP-SVD	1.4	**8.3B**	6.41	11.10	**41.13**	68.60	**70.40**	69.85	66.77	41.60	73.67	34.70	58.34
	+NMP-SVD	1.5	8.9B	**6.03**	**9.78**	40.87	**69.28**	**70.40**	**76.18**	**70.07**	**43.60**	**75.35**	**36.21**	**60.24**
	KFAC-SVD	–	10.5B	6.67	13.07	40.36	69.82	69.38	66.02	63.86	40.20	73.23	38.38	57.66
	+NMP-SVD	1.4	**8.3B**	6.38	11.06	40.36	68.94	69.22	64.86	66.51	41.40	73.94	37.81	57.88
	+NMP-SVD	1.5	8.9B	**5.98**	**9.76**	**43.26**	**71.93**	**70.80**	**71.16**	**69.69**	**42.20**	**75.19**	**40.44**	**60.58**
40%	ASVD	–	7.9B	55.82	75.75	29.78	44.57	60.14	69.42	44.12	33.00	63.38	25.64	46.26
	+NMP-SVD	1.4	**6.3B**	14.61	22.69	34.30	55.98	64.72	69.08	57.11	37.40	69.70	28.24	52.07
	+NMP-SVD	1.5	6.7B	**11.05**	**16.61**	**35.49**	**58.04**	**65.04**	**76.51**	**62.77**	**40.20**	**73.50**	**31.72**	**55.41**
	SVD-LLM	–	7.9B	11.79	43.62	28.50	48.02	62.19	59.82	43.14	34.00	61.32	24.91	45.24
	+NMP-SVD	1.4	**6.3B**	10.06	29.23	29.69	50.63	63.14	62.75	46.72	37.00	64.42	23.50	47.23
	+NMP-SVD	1.5	6.7B	**8.79**	**22.25**	**32.68**	**55.89**	**65.75**	**64.43**	**51.37**	**37.40**	**66.76**	**25.37**	**49.96**
	KFAC-SVD	–	7.9B	11.58	37.10	29.61	50.80	61.80	59.05	44.38	35.60	62.02	25.55	46.10
	+NMP-SVD	1.4	**6.3B**	10.06	27.28	30.38	54.34	63.54	**64.10**	47.67	36.40	64.64	25.00	48.26
	+NMP-SVD	1.5	6.7B	**8.73**	**20.50**	**32.59**	**58.88**	**65.04**	62.78	**52.94**	**37.20**	**68.44**	**26.36**	**50.53**
60%	ASVD	–	5.4B	1.7e4	1.7e4	**26.11**	27.06	51.22	53.00	27.49	22.00	51.47	**23.52**	35.23
	+NMP-SVD	1.4	**4.3B**	1992	3059	25.85	29.63	52.57	**55.47**	29.11	24.60	51.14	23.42	36.47
	+NMP-SVD	1.5	4.6B	**596.8**	**1204**	25.85	**30.22**	**54.06**	54.95	**31.24**	**25.80**	**53.16**	23.14	**37.30**
	SVD-LLM	–	5.4B	44.93	257.5	**24.91**	31.02	52.72	**37.83**	29.31	26.80	52.61	**23.86**	34.88
	+NMP-SVD	1.4	**4.3B**	36.52	194.9	23.55	33.46	52.25	**37.83**	30.66	26.80	53.81	23.77	35.27
	+NMP-SVD	1.5	4.6B	**28.55**	**143.4**	23.29	**34.09**	**55.88**	**37.83**	**31.92**	**27.40**	**54.41**	23.14	**35.99**
	KFAC-SVD	–	5.4B	43.47	210.5	23.89	31.10	**53.99**	37.83	30.23	26.00	51.74	**23.00**	34.72
	+NMP-SVD	1.4	**4.3B**	34.80	165.7	**24.40**	31.99	51.93	**37.89**	31.37	**29.40**	52.83	22.98	35.35
	+NMP-SVD	1.5	4.6B	**28.13**	**125.2**	23.38	**32.95**	53.67	37.83	**32.21**	28.00	**53.92**	22.98	**35.62**

F.2 Detail Results under Uniform Rank Allocation with Adaptive Dense-Sparse Selection

Detailed evaluation results for Llama2-70B is shown in Table 13. We report equivalent parameter count (P_{eq}), perplexity on WikiText2 (WT2) and C4, and zero-shot accuracy (%) on ARC-c, ARC-e, WinoGrande (WG), BoolQ (BQ), HellaSwag (HS), OpenBookQA (OBQA), PIQA, MMLU, and their average (AVG). The best result within each baseline group for each metric is highlighted in bold.

F.3 Detail Results under Non-uniform Rank Allocation with Adaptive Dense-Sparse Selection

Detailed evaluation results for Llama3-8B and Qwen3-8B are shown in Table 14 and Table 15, respectively. We report equivalent parameter count (P_{eq}), perplexity on WikiText2 (WT2) and C4, and zero-shot accuracy (%) on ARC-c, ARC-e, WinoGrande (WG), BoolQ (BQ), HellaSwag (HS), OpenBookQA (OBQA), PIQA, MMLU, and their average (AVG). The best result within each baseline group for each metric is highlighted in bold.

Table 10: Detailed evaluation results for Llama3-8B.

Ratio	Method	ρ	$P_{\text{eq}}\downarrow$	WT2↓	C4↓	ARC-c↑	ARC-e↑	WG↑	BQ↑	HS↑	OBQA↑	PIQA↑	MMLU↑	**AVG↑**
0%	Dense	–	8B	6.13	9.26	54.10	77.53	73.80	82.26	79.23	44.80	80.63	62.36	69.34
20%	ASVD	–	6.6B	521.3	471.3	25.17	35.48	54.70	50.15	30.51	26.60	57.89	23.30	37.98
	+NMP-SVD	1.4	**5.4B**	16.85	26.14	39.42	64.14	67.96	**76.82**	64.86	41.80	73.18	41.09	58.66
	+NMP-SVD	1.5	5.8B	**10.98**	**17.11**	**42.83**	**67.09**	**68.75**	70.55	**70.75**	**42.80**	**75.84**	**45.46**	**60.51**
	SVD-LLM	–	6.6B	15.60	79.71	29.27	50.21	63.61	63.12	41.50	32.40	60.45	23.37	45.49
	+NMP-SVD	1.4	**5.4B**	12.73	35.90	32.68	57.49	64.17	66.02	53.89	37.00	70.13	26.96	51.04
	+NMP-SVD	1.5	5.8B	**10.48**	**26.21**	**36.77**	**61.74**	**67.48**	**71.01**	**59.57**	**38.40**	**73.56**	**30.74**	**54.91**
	KFAC-SVD	–	6.6B	13.67	61.78	33.53	54.08	64.01	62.20	49.09	33.60	65.56	23.96	48.25
	+NMP-SVD	1.4	**5.4B**	12.49	34.54	38.82	64.31	66.46	62.48	56.00	37.00	69.48	35.12	53.71
	+NMP-SVD	1.5	5.8B	**10.40**	**23.39**	**38.99**	**67.26**	**67.64**	**64.04**	**62.00**	**37.20**	**72.25**	**37.36**	**55.84**
40%	ASVD	–	5.2B	1.0e5	1.6e5	24.23	25.88	50.51	42.35	25.61	26.40	49.73	**26.34**	33.88
	+NMP-SVD	1.4	**4.3B**	1958	1404	24.23	31.27	53.43	53.58	30.13	27.20	54.24	23.27	37.17
	+NMP-SVD	1.5	4.6B	**683.0**	**554.0**	**26.45**	**35.90**	**55.09**	**61.50**	**36.81**	**30.80**	**59.14**	23.55	**41.15**
	SVD-LLM	–	5.2B	87.76	836.4	**24.83**	29.80	52.72	37.83	26.84	27.40	52.07	23.72	34.40
	+NMP-SVD	1.4	**4.3B**	56.42	392.6	22.53	31.73	53.12	39.57	28.25	27.80	54.30	23.61	35.11
	+NMP-SVD	1.5	4.6B	**32.10**	**174.2**	24.23	**37.08**	**57.46**	**57.37**	**32.89**	**29.40**	**57.34**	**25.09**	**40.11**
	KFAC-SVD	–	5.2B	65.00	563.0	23.29	33.04	53.04	37.83	28.50	29.20	53.10	**24.21**	35.28
	+NMP-SVD	1.4	**4.3B**	50.67	371.8	24.32	35.86	54.54	41.04	31.55	30.00	55.71	23.79	37.10
	+NMP-SVD	1.5	4.6B	**28.96**	**157.9**	**26.37**	**45.58**	**58.48**	**61.07**	**37.09**	**31.20**	**60.83**	23.00	**42.95**
60%	ASVD	–	3.8B	7.5e5	8.8e5	26.28	24.66	49.88	**61.19**	**26.48**	**29.20**	**52.29**	**26.89**	**37.11**
	+NMP-SVD	1.4	**3.3B**	2.9e5	2.8e5	25.00	**27.23**	**51.30**	55.11	26.14	26.40	50.54	25.54	35.91
	+NMP-SVD	1.5	3.4B	**4.9e4**	**4.2e4**	**26.62**	25.00	50.20	46.64	26.29	25.80	51.41	23.89	34.48
	SVD-LLM	–	3.8B	819.8	3197	**25.00**	26.26	50.83	**37.83**	25.77	23.00	50.16	**24.07**	32.87
	+NMP-SVD	1.4	**3.3B**	901.7	3963	24.83	26.64	**53.59**	**37.83**	25.73	26.20	49.73	23.84	**33.55**
	+NMP-SVD	1.5	3.4B	**748.5**	**3016**	24.23	**27.15**	50.75	**37.83**	**26.06**	**26.80**	**50.33**	22.99	33.27
	KFAC-SVD	–	3.8B	651.4	3734	24.83	25.46	**50.75**	**37.89**	26.16	26.20	**51.36**	24.41	33.38
	+NMP-SVD	1.4	**3.3B**	598.7	2760	**26.62**	27.86	**50.75**	37.83	26.09	**26.40**	49.73	**24.43**	**33.71**
	+NMP-SVD	1.5	3.4B	**492.2**	**2668**	25.17	**28.16**	48.30	37.83	**26.37**	26.00	48.75	24.33	33.11

Table 11: Detailed evaluation results for Qwen3-8B.

Ratio	Method	ρ	$P_{\text{eq}}\downarrow$	WT2↓	C4↓	ARC-c↑	ARC-e↑	WG↑	BQ↑	HS↑	OBQA↑	PIQA↑	MMLU↑	**AVG↑**
0%	Dense	–	8.2B	9.71	15.23	56.48	80.93	67.80	86.57	74.93	41.40	77.80	72.97	69.86
20%	ASVD	–	6.8B	30.83	40.68	35.84	50.93	59.27	67.34	57.01	36.20	68.99	44.70	52.53
	+NMP-SVD	1.4	**5.6B**	15.18	25.17	43.86	66.08	63.77	78.56	65.32	40.20	73.67	58.14	61.20
	+NMP-SVD	1.5	5.9B	**13.33**	**22.12**	**50.00**	**72.10**	**65.27**	**82.57**	**67.61**	**42.20**	**74.16**	**60.47**	**64.30**
	SVD-LLM	–	6.8B	15.07	36.58	40.19	60.52	64.09	79.20	59.45	39.00	69.97	48.94	57.67
	+NMP-SVD	1.4	**5.6B**	13.36	27.90	45.14	**67.21**	**67.25**	83.70	64.21	**40.00**	73.07	54.00	61.82
	+NMP-SVD	1.5	5.9B	**12.63**	**24.19**	**45.31**	63.43	65.82	**84.19**	**66.86**	39.80	**73.39**	**60.12**	**62.37**
	KFAC-SVD	–	6.8B	13.50	31.14	43.17	63.26	65.67	75.57	59.37	38.00	69.59	52.64	58.41
	+NMP-SVD	1.4	**5.6B**	12.98	26.12	**46.76**	**68.52**	**66.77**	80.92	62.06	37.80	72.09	56.87	61.47
	+NMP-SVD	1.5	5.9B	**11.90**	**22.27**	45.05	67.42	65.90	**85.11**	**65.54**	**41.20**	**74.05**	**60.74**	**63.13**
40%	ASVD	–	5.4B	251.5	248.1	27.99	40.40	53.43	51.01	37.57	30.60	59.47	27.00	40.93
	+NMP-SVD	1.4	**4.5B**	48.92	80.47	31.06	46.84	57.30	65.47	47.77	32.80	64.25	30.86	47.04
	+NMP-SVD	1.5	4.8B	**32.78**	**53.35**	**33.53**	**54.29**	**59.51**	**71.19**	**53.40**	**35.20**	**66.97**	**38.61**	**51.59**
	SVD-LLM	–	5.4B	28.00	113.8	25.51	36.62	54.62	43.70	36.76	28.20	58.60	22.98	38.37
	+NMP-SVD	1.4	**4.5B**	21.94	91.03	33.02	50.67	59.67	**74.04**	46.04	33.20	63.60	29.50	48.72
	+NMP-SVD	1.5	4.8B	**19.22**	**61.80**	**36.77**	**53.83**	**61.56**	71.56	**51.63**	**34.60**	**67.41**	**37.72**	**51.89**
	KFAC-SVD	–	5.4B	23.23	100.8	27.05	40.03	58.56	54.62	39.26	31.80	60.12	24.77	42.02
	+NMP-SVD	1.4	**4.5B**	19.71	69.18	30.55	**54.29**	60.69	64.50	44.73	33.20	63.93	30.16	47.76
	+NMP-SVD	1.5	4.8B	**17.31**	**52.97**	**33.96**	53.37	**62.90**	**76.02**	**49.45**	**34.20**	**65.67**	**40.39**	**52.00**
60%	ASVD	–	4B	2.8e4	8.4e4	**26.88**	28.20	51.38	42.81	27.42	27.40	53.32	24.10	35.19
	+NMP-SVD	1.4	**3.4B**	2269	4053	23.38	30.72	49.49	**51.10**	29.89	**29.20**	54.03	24.61	36.55
	+NMP-SVD	1.5	3.6B	**815.9**	**2119**	25.34	**30.89**	**51.46**	48.56	**31.96**	27.40	**55.82**	**25.23**	**37.08**
	SVD-LLM	–	4B	182.0	1508	**22.70**	27.95	49.41	**38.47**	28.83	25.00	52.12	**23.12**	33.45
	+NMP-SVD	1.4	**3.4B**	142.1	656.3	**22.70**	27.57	**50.59**	37.80	29.84	24.60	54.03	23.01	33.77
	+NMP-SVD	1.5	3.6B	**115.7**	**476.8**	22.10	**29.17**	49.57	37.86	**30.83**	**28.00**	**54.30**	23.02	**34.35**
	KFAC-SVD	–	4B	126.1	541.4	21.84	27.90	52.17	**37.83**	29.11	26.40	52.67	22.86	33.85
	+NMP-SVD	1.4	**3.4B**	103.1	392.7	**22.87**	30.01	51.14	**37.83**	30.33	26.40	53.21	**23.16**	34.37
	+NMP-SVD	1.5	3.6B	**97.73**	**378.3**	21.84	**31.48**	**53.75**	**37.83**	**31.43**	**28.80**	**54.79**	23.07	**35.37**

Table 12: Detailed evaluation results for Qwen3-32B.

Ratio	Method	ρ	$P_{\text{eq}}\downarrow$	WT2↓	C4↓	ARC-c↑	ARC-e↑	WG↑	BQ↑	HS↑	OBQA↑	PIQA↑	MMLU↑	**AVG↑**
0%	Dense	–	32.8B	7.61	12.23	60.84	83.25	80.81	86.57	82.62	82.21	73.16	46.60	74.51
20%	ASVD	–	26.5B	18.88	29.36	44.45	64.98	66.99	84.25	68.85	73.18	64.17	41.00	63.48
	+NMP-SVD	1.4	**21.2B**	11.26	17.58	52.05	72.85	71.79	86.39	77.49	77.80	66.61	42.40	68.42
	+NMP-SVD	1.5	22.6B	**9.22**	**15.02**	**57.42**	**80.60**	**74.73**	**87.61**	**78.97**	**80.03**	**68.35**	**44.00**	**71.46**
	SVD-LLM	–	26.5B	11.09	27.70	51.02	70.58	68.07	**86.88**	71.75	74.21	69.30	43.80	66.95
	+NMP-SVD	1.4	**21.2B**	9.99	18.40	**56.14**	**77.27**	69.28	79.51	75.93	77.80	**70.64**	**45.80**	69.05
	+NMP-SVD	1.5	22.6B	**9.35**	**15.76**	55.38	74.96	**72.53**	86.30	**77.41**	**78.13**	70.24	44.20	**69.89**
	KFAC-SVD	–	26.5B	9.92	22.11	52.05	69.19	69.71	84.62	71.95	75.84	69.38	43.80	67.07
	+NMP-SVD	1.4	**21.2B**	9.34	17.03	**53.92**	**73.15**	72.08	86.79	75.10	76.88	71.35	**46.00**	69.41
	+NMP-SVD	1.5	22.6B	**8.86**	**15.36**	53.16	71.21	**74.63**	**88.41**	**77.10**	**78.24**	**72.69**	45.80	**70.15**
40%	ASVD	–	20.3B	68.65	99.27	36.18	51.26	42.62	75.26	51.67	65.18	56.99	35.20	51.79
	+NMP-SVD	1.4	**16.3B**	25.63	41.51	40.70	57.70	49.62	81.47	62.99	69.15	61.64	38.20	57.68
	+NMP-SVD	1.5	17.4B	**19.55**	**31.90**	**46.33**	**63.01**	**59.23**	**82.84**	**67.54**	**71.44**	**62.04**	**40.60**	**61.63**
	SVD-LLM	–	20.3B	17.50	81.38	37.88	58.80	41.37	**79.11**	48.27	64.15	63.22	36.60	53.67
	+NMP-SVD	1.4	**16.3B**	14.51	49.70	45.31	67.76	51.70	72.84	59.36	69.97	68.59	**40.40**	59.49
	+NMP-SVD	1.5	17.4B	**12.69**	**35.65**	**47.01**	**71.00**	**59.09**	74.46	**64.38**	**72.25**	**69.30**	39.60	**62.14**
	KFAC-SVD	–	20.3B	14.65	56.85	41.47	61.20	44.94	**74.86**	51.57	67.52	65.11	37.60	55.53
	+NMP-SVD	1.4	**16.3B**	12.85	38.28	44.71	**64.23**	57.83	69.63	59.30	70.95	65.67	41.20	59.19
	+NMP-SVD	1.5	17.4B	**11.70**	**30.45**	**48.21**	63.47	**62.61**	74.10	**64.09**	**72.58**	**68.11**	**43.40**	**62.07**
60%	ASVD	–	14B	1311	2452	25.85	30.26	24.62	59.27	29.46	54.95	49.96	29.00	37.92
	+NMP-SVD	1.4	**11.4B**	253.3	447.5	27.05	36.11	26.43	59.60	37.61	57.18	51.70	30.20	40.73
	+NMP-SVD	1.5	12.1B	**142.9**	**233.9**	**27.82**	**38.13**	**26.71**	**62.48**	**41.45**	**58.05**	**54.22**	**32.00**	**42.61**
	SVD-LLM	–	14B	72.77	460.7	23.38	31.65	23.42	37.83	30.91	54.35	52.72	24.60	34.86
	+NMP-SVD	1.4	**11.4B**	45.55	315.2	23.55	35.86	24.66	37.92	33.56	55.88	52.17	**27.40**	36.37
	+NMP-SVD	1.5	12.1B	**33.58**	**196.5**	**26.11**	**38.64**	**26.27**	**38.81**	**36.17**	**58.11**	**56.91**	27.00	**38.50**
	KFAC-SVD	–	14B	46.85	269.2	23.72	33.54	23.76	38.10	32.14	55.06	53.91	28.20	36.05
	+NMP-SVD	1.4	**11.4B**	30.26	169.5	26.96	37.21	26.24	40.15	35.62	57.67	56.59	**31.00**	38.93
	+NMP-SVD	1.5	12.1B	**24.49**	**132.1**	**28.41**	**43.86**	**26.43**	**62.81**	**38.32**	**58.87**	**56.99**	30.20	**43.24**

Table 13: Detailed evaluation results for Llama2-70B.

Ratio	Method	ρ	$P_{\text{eq}}\downarrow$	WT2↓	C4↓	ARC-c↑	ARC-e↑	WG↑	BQ↑	HS↑	OBQA↑	PIQA↑	MMLU↑	**AVG↑**
0%	Dense	–	70B	3.32	5.64	57.17	79.50	80.43	85.14	84.34	49.40	82.59	65.61	73.02
20%	SVD-LLM	–	55.3B	4.72	8.49	54.35	81.31	77.58	67.22	75.55	46.60	78.45	55.85	67.12
	+NMP-SVD	1.4	**43.7B**	4.49	7.41	**57.94**	**81.90**	77.51	75.63	78.90	46.80	79.22	57.38	69.41
	+NMP-SVD	1.5	46.7B	**4.26**	**6.92**	55.29	79.71	**78.61**	**79.45**	**80.08**	**47.40**	**79.92**	**58.19**	**69.83**
	KFAC-SVD	–	55.3B	4.69	8.13	52.90	79.34	77.35	65.35	75.86	45.80	78.89	57.29	66.60
	+NMP-SVD	1.4	**43.7B**	4.64	7.82	54.01	**80.51**	76.64	69.82	77.09	44.80	79.60	56.09	67.32
	+NMP-SVD	1.5	49B	**4.33**	**7.12**	**54.69**	79.76	**78.14**	**73.55**	**78.90**	**47.60**	**79.82**	**59.29**	**68.97**
40%	SVD-LLM	–	41.6B	7.05	20.55	43.09	71.55	72.93	62.42	54.76	40.80	69.26	34.97	56.22
	+NMP-SVD	1.4	**33B**	6.29	13.70	47.95	76.94	74.27	63.12	64.42	41.40	73.12	41.27	60.31
	+NMP-SVD	1.5	35.2B	**5.66**	**11.19**	**50.94**	**78.41**	**76.01**	**66.33**	**69.03**	**44.80**	**76.71**	**46.75**	**63.62**
	KFAC-SVD	–	41.6B	6.89	17.36	43.94	72.81	72.85	**63.06**	54.47	39.80	70.08	39.67	57.09
	+NMP-SVD	1.4	**33B**	6.57	14.08	45.65	73.61	73.09	62.29	61.20	39.40	71.44	42.13	58.60
	+NMP-SVD	1.5	36.6B	**5.89**	**11.76**	**47.53**	**76.94**	**76.16**	62.66	**66.09**	**42.00**	**74.65**	**47.37**	**61.67**
60%	SVD-LLM	–	27.9B	17.13	97.94	24.57	38.76	57.30	37.92	33.86	29.40	55.39	23.19	37.55
	+NMP-SVD	1.4	**22.2B**	13.86	70.12	26.96	44.02	59.59	48.32	36.32	32.20	58.38	23.54	41.17
	+NMP-SVD	1.5	24B	**11.68**	**52.93**	**29.52**	**49.75**	**64.17**	**62.02**	**40.11**	**34.60**	**60.88**	**24.35**	**45.67**
	KFAC-SVD	–	27.9B	18.02	86.67	24.15	39.73	55.88	37.83	33.38	29.60	56.31	22.87	37.47
	+NMP-SVD	1.4	**22.2B**	14.78	60.76	24.32	45.83	57.70	52.45	35.42	**31.00**	57.45	23.09	40.91
	+NMP-SVD	1.5	24.4B	**12.56**	**47.14**	**26.79**	**49.58**	**59.75**	**61.41**	**37.14**	**31.00**	**59.90**	**23.15**	**43.59**

Table 14: Detailed non-uniform compression results for Llama3-8B.

Ratio	**Method**	ρ	$P_{\text{eq}}\downarrow$	WT2↓	C4↓	ARC-c↑	ARC-e↑	WG↑	BQ↑	HS↑	OBQA↑	PIQA↑	MMLU↑	**AVG↑**
0%	Dense	–	8B	6.13	9.26	54.10	77.53	73.80	82.26	79.23	44.80	80.63	62.36	69.34
20%	Zero-Sum SVD	–	6.6B	12.51	33.58	37.20	64.60	59.35	63.58	54.18	37.80	71.33	26.03	51.76
	+NMP-SVD	1.4	**6.2B**	11.78	23.96	37.03	63.93	61.64	71.28	58.82	38.80	74.21	**29.06**	54.35
	+NMP-SVD	1.5	6.3B	**10.78**	**20.55**	**40.87**	**68.43**	**64.48**	**72.57**	**62.28**	**42.00**	**75.14**	27.97	**56.72**
	SVD-LLM-v2	–	6.6B	15.98	90.93	28.50	48.27	62.98	61.10	39.45	31.40	60.12	22.87	44.34
	+NMP-SVD	1.4	**5.6B**	12.52	37.09	36.09	58.46	66.93	71.80	53.75	38.20	68.55	27.20	52.62
	+NMP-SVD	1.5	5.8B	**10.43**	**26.23**	**39.08**	**63.93**	**68.98**	**75.72**	**61.88**	**42.20**	**73.45**	**33.55**	**57.35**
	KFAC-LEMS	–	6.6B	9.33	17.46	44.03	71.34	68.43	63.43	68.15	**43.20**	76.28	51.92	60.85
	+NMP-SVD	1.4	**6.3B**	8.84	16.02	47.70	75.80	69.30	65.14	69.04	41.80	**76.71**	52.65	62.27
	+NMP-SVD	1.5	6.4B	**8.45**	**14.98**	**48.04**	**77.36**	**70.32**	**65.29**	**70.83**	42.80	76.55	**53.70**	**63.11**
40%	Zero-Sum SVD	–	5.2B	**52.16**	455.6	**26.37**	26.47	50.67	**37.83**	25.19	**26.80**	50.92	24.13	33.55
	+NMP-SVD	1.4	**4.5B**	116.3	564.3	22.70	30.26	50.04	**37.83**	26.84	**26.80**	52.23	25.40	34.01
	+NMP-SVD	1.5	4.7B	79.87	**445.2**	23.21	**31.14**	**50.99**	**37.83**	**28.28**	26.60	**52.88**	**26.01**	**34.62**
	SVD-LLM-v2	–	5.2B	114.8	1117	**24.40**	28.16	52.49	37.83	27.01	26.40	50.82	**25.20**	34.04
	+NMP-SVD	1.4	**4.4B**	79.36	510.6	23.89	28.28	51.70	38.01	27.18	26.80	51.58	24.63	34.01
	+NMP-SVD	1.5	4.6B	**46.19**	**287.3**	23.46	**32.28**	**54.06**	**44.46**	**29.45**	**29.00**	**55.06**	23.48	**36.41**
	KFAC-LEMS	–	5.2B	25.49	97.69	27.30	45.88	60.14	49.20	43.16	31.40	63.11	24.47	43.08
	+NMP-SVD	1.4	**4.8B**	21.72	69.31	28.84	48.78	61.40	62.11	46.45	32.60	66.16	25.22	46.45
	+NMP-SVD	1.5	4.9B	**18.44**	**52.44**	**29.35**	**52.23**	**62.12**	**62.20**	**49.76**	**33.40**	**66.92**	**30.50**	**48.31**

Table 15: Detailed non-uniform compression results for Qwen3-8B.

Ratio	**Method**	ρ	$P_{\text{eq}}\downarrow$	WT2↓	C4↓	ARC-c↑	ARC-e↑	WG↑	BQ↑	HS↑	OBQA↑	PIQA↑	MMLU↑	**AVG↑**
0%	Dense	–	8.2B	9.71	15.23	56.48	80.93	67.80	86.57	74.93	41.40	77.80	72.97	69.86
20%	Zero-Sum SVD	–	6.8B	**11.45**	23.41	52.47	**80.09**	**66.69**	84.28	67.82	40.00	76.06	59.88	65.91
	+NMP-SVD	1.4	**6.2B**	11.69	20.48	51.45	78.24	66.06	**84.59**	67.93	**42.20**	**76.71**	61.28	66.06
	+NMP-SVD	1.5	6.4B	11.53	**19.21**	**53.92**	78.49	66.06	84.37	**69.22**	**42.20**	76.55	**62.81**	**66.70**
	SVD-LLM-v2	–	6.8B	14.94	37.22	40.10	56.78	64.33	78.87	59.70	38.80	70.18	47.70	57.06
	+NMP-SVD	1.4	**5.9B**	13.40	29.63	45.31	62.08	65.90	83.49	64.85	**40.20**	73.01	55.35	61.27
	+NMP-SVD	1.5	6B	**12.41**	**22.28**	**47.53**	**64.69**	**67.96**	**84.10**	**66.42**	39.00	**73.07**	**58.15**	**62.61**
	KFAC-LEMS	–	6.8B	11.30	19.71	50.51	**77.61**	69.53	84.95	66.01	40.60	74.81	62.54	65.82
	+NMP-SVD	1.4	**6.4B**	10.99	18.55	50.77	76.35	69.38	85.96	66.63	40.60	75.84	65.00	66.32
	+NMP-SVD	1.5	6.5B	**10.72**	**17.82**	**52.47**	76.94	**70.64**	**86.15**	**67.78**	**41.20**	**76.01**	**66.35**	**67.19**
40%	Zero-Sum SVD	–	5.4B	23.21	95.50	33.96	50.51	**59.27**	61.96	45.68	**34.00**	66.49	28.17	47.50
	+NMP-SVD	1.4	**4.6B**	22.52	84.69	33.45	52.36	55.96	66.09	44.78	32.00	65.56	28.93	47.39
	+NMP-SVD	1.5	4.8B	**19.99**	**68.02**	**36.26**	**59.30**	59.04	**73.03**	**48.13**	32.40	**67.08**	**33.96**	**51.15**
	SVD-LLM-v2	–	5.4B	32.31	140.6	24.66	36.41	54.22	40.95	35.30	29.00	56.69	22.99	37.53
	+NMP-SVD	1.4	**4.6B**	22.50	86.84	30.80	49.24	60.22	71.68	45.27	32.20	64.74	30.46	48.08
	+NMP-SVD	1.5	4.8B	**19.81**	**68.11**	**35.49**	**51.09**	**61.33**	**75.57**	**49.92**	**34.00**	**66.54**	**34.08**	**51.00**
	KFAC-LEMS	–	5.4B	20.59	62.35	32.51	55.85	62.19	55.32	46.11	31.60	65.67	32.88	47.77
	+NMP-SVD	1.4	**4.9B**	20.89	55.10	34.64	56.31	62.75	70.80	48.03	32.60	67.08	34.80	50.88
	+NMP-SVD	1.5	5.1B	**18.32**	**47.47**	**35.75**	**57.79**	**63.14**	**76.39**	**50.22**	**32.80**	**67.14**	**37.71**	**52.62**