LEARNING STRUCTURED SPARSITY FOR LARGE MODELS: BRIDGING MODEL COMPRESSIBILITY AND EFFICIENT INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling laws continue to reward increases in model scale, while the resulting computational and memory demands drive greater reliance on cloud infrastructure for inference. Yet unreliable network connectivity, communication delays, and privacy requirements can make such reliance impractical, motivating local execution on edge devices. Local deployment requires models to fit within limited memory and execute efficiently with the available computational resources. Recent work has shown that learning sparse-plus-low-rank representations enables elastic compression of a single checkpoint to multiple memory budgets. However, fewer parameters do not necessarily translate into faster inference. Indexing overhead and irregular memory access in unstructured sparse components can offset computational savings. To address this critical deployment barrier, we introduce S-SALAAD, a unified framework that jointly learns the model weights and their low-rank and structured sparse components during training. The framework supports a family of hardware-compatible structured sparsity patterns. It couples this structure learning with budget-aware compression and dedicated CPU and GPU kernels that exploit the learned structure for acceleration, enabling efficient inference from a single checkpoint across memory budgets. Experiments on Llama2 and Qwen3 models, covering both pretraining and distillation, show that S-SALAAD advances the empirical Pareto frontier in perplexity, memory, and inference latency across multiple hardware platforms relative to the evaluated baselines. These results show that structured learning can translate elastic compressibility into reduced memory requirements and lower inference latency, enabling practical deployment of large models on edge devices.

1 INTRODUCTION

Scaling laws (Kaplan et al., 2020) motivate larger models for improved quality, while their memory and computational demands increase reliance on cloud infrastructure with high-end accelerators. However, network connectivity can be unstable in underwater and rescue operations (Stojanovic & Preisig, 2009; Deng et al., 2013; Gunn & Anderson, 2015). Safety-critical applications such as autonomous driving require on-device decisions to avoid cloud communication delays (Liu et al., 2019; Kuutti et al., 2021; Ma et al., 2025; Bodmer et al., 2026), while privacy requirements can prevent sensitive data from being uploaded (European Parliament and Council of the European Union, 2016; Shi et al., 2016). These constraints require large-model inference to move from the cloud to resource-constrained edge devices. Fitting large models within limited memory and achieving low inference latency with the available computational resources are the two critical challenges of this transition. The latency requirement is particularly important in feedback control, where inference latency adds to loop delay and excessive delays can degrade closed-loop performance or destabilize the system, compromising safety (Shin & Cui, 1995; Findeisen & Allgöwer, 2004; Ma et al., 2026b; 2023; Nan et al., 2026).

To address memory constraints, existing compression methods use quantization (Dettmers et al., 2022), distillation (Gu et al., 2024), pruning (Sun et al., 2024; Frantar & Alistarh, 2023; Ashkboos et al., 2024), and low-rank factorization (Hsu et al., 2022; Wang et al., 2025c). Varying memory budgets across edge devices call for elastic compression of a single trained checkpoint to multiple

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.

model sizes. SALAAD (Ma et al., 2026a) enables such elastic compression by learning sparse-and-low-rank (SLR) representations for each linear projection[1] during pretraining. *However, fewer parameters do not necessarily translate into lower inference latency.* Unstructured SLR saves static memory but incurs higher inference latency (see Appendix A.1). We identify the overhead from irregular memory access in unstructured sparse operations as a critical barrier to deploying large models on edge devices. Hardware-friendly patterns, including block sparsity (Okanovic et al., 2025), $N:M$ sparsity (Zhao et al., 2024), and row-/column-wise sparsity (Wen et al., 2016), are therefore required to translate these memory savings into efficient edge inference. A straightforward extension of SALAAD is to impose these patterns on its learned sparse components after pretraining. However, our experiments show that this post-hoc conversion can sharply degrade model performance (see Appendix A.2). We therefore induce *structured sparsity* during *pretraining*, allowing the SLR components to adapt jointly to the structural constraints.

To this end, we introduce **S-SALAAD** (Structured-Sparse-And-Low-Rank Adaptation via ADMM), a plug-and-play framework that learns *structured* SLR representations during pretraining for elastic compression and efficient edge inference. Based on the alternating direction method of multipliers (ADMM), it induces block, $N:M$, and row-/column-wise sparsity without changing the model architecture or optimizer update rules. Efficient low-rank updates reduce pretraining overhead, while budget-aware refinement enables zero-shot compression of a single checkpoint to multiple parameter budgets without calibration data or further training. Among these patterns, we identify block sparsity as a practical choice for elastic deployment given current technical capabilities. With dedicated CPU and GPU kernels, the resulting models advance the perplexity–memory–latency Pareto frontier relative to the evaluated baselines. Our contributions are summarized as follows:

- **A plug-and-play framework for structured-sparsity learning.** S-SALAAD jointly learns model weights and their low-rank and structured sparse components under the task loss, adapting the representation to structural constraints during pretraining. S-SALAAD induces block, $N:M$, and row-/column-wise sparsity without changing the model architecture, and it is compatible with existing optimizers.
- **A practical end-to-end implementation.** To support the training of larger models and their edge deployment, we address bottlenecks in the training, compression, and inference stages. Specifically, we use i) a warm-started iterative singular value decomposition (I-SVD) to reduce the computational overhead of low-rank updates during pretraining; ii) energy-guided alternating refinement (EGAR) to enable zero-shot compression without calibration data or further training, allocating the parameter budget across projections without a preset sparse–low-rank split and reducing reconstruction error within the allocated capacities; and iii) dedicated CPU and GPU kernels to exploit the induced structured sparsity and reduce inference latency.
- **Empirical Pareto frontier across training settings.** S-SALAAD advances the empirical perplexity–memory–latency Pareto frontier across hardware platforms relative to the evaluated baselines, with each checkpoint supporting multiple compression budgets. These gains hold for Llama2-1B pretraining and cross-architecture distillation from Nemotron-3-Super-120B to Qwen3-1.7B.
- **Analysis of structured representations.** S-SALAAD provides an explicit structured representation of Transformer weights, enabling us to identify distinct sparsity characteristics across linear projection types. Our analysis identifies block sparsity as a practical choice for advancing the perplexity-memory-latency Pareto frontier, combining flexible compression with efficient execution on CPUs and GPUs. Furthermore, we also reveal the relationship between tolerable block granularity and model scale, providing empirical guidance for selecting block sizes when extending S-SALAAD to larger models.

2 Related Work

Compression methods for large language models broadly divide into pruning, low-rank factorization, and quantization. Unstructured pruning removes individual weights by magnitude (Han et al., 2015) or calibration-based criteria (Sun et al., 2024; Frantar & Alistarh, 2023), while low-rank

[1]Here, *linear projection* refers to a learned linear map in the attention (query/key/value/output) or multilayer perceptron (MLP; up/gate/down) modules. Embeddings and the language model head are excluded unless noted.

methods such as SVD-LLM (Wang et al., 2025c) factorize weight matrices and SliceGPT (Ashkboos et al., 2024) shrinks them by removing channels. A second post-hoc family instead runs iterative ADMM on the pretrained weights: ELSA (Lee et al., 2026) pushes unstructured sparsity to extreme ratios, and 3BASiL (Makni et al., 2025) performs layer-wise SLR reconstruction via a 3-block ADMM–both are data-hungry and take hours of GPU-time to compress a single model. Compression-aware pretraining instead induces compressibility during training: LORO (Mo et al., 2025) pretrains in a preset low-rank subspace, SLTrain (Han et al., 2024) augments low-rank pretraining with a fixed-support sparse component, and SALAAD (Ma et al., 2026a) jointly learns SLR components via ADMM. The sparse components in SLTrain and SALAAD are unstructured and do not map efficiently onto accelerator primitives. LOST (Li et al., 2025) combines low-rank factors with channel-wise sparsity, but its preset rank and sparse capacity fix the model size during training, and its formulation does not support elastic deployment across budgets from a single checkpoint. It also inserts an additional nonlinearity between the low-rank factors, changing the underlying model architecture. S-SALAAD preserves the original architecture and supports multiple structured sparsity patterns, learning a representation that can be compressed from a single checkpoint to different parameter budgets.

3 Problem Formulation

Let $\{X_i\}_{i=1}^{E}$ denote the weight matrices of a neural network's E target linear projections, where $X_i \in \mathbb{R}^{n_i \times m_i}$. For notational simplicity, we omit the projection index i below. SALAAD (Ma et al., 2026a) decomposes each projection matrix X into a low-rank component $L \in \mathbb{R}^{n\times m}$ and an unstructured element-wise sparse component $S \in \mathbb{R}^{n\times m}$ by solving the following optimization problem:

$$\min_{X,L,S} \ \ell(X) + \alpha \left|L\right|_* + \beta \left|S\right|_1 \quad \text{s.t. } X = L + S, \tag{1}$$

where $\ell : \mathbb{R}^{n\times m} \to \mathbb{R}$ denotes the task-specific loss. The nuclear norm $\left|\cdot\right|_*$ and ℓ_1-norm $\left|\cdot\right|_1$ serve as convex surrogates for matrix rank and the number of nonzero entries, respectively, while $\alpha > 0$ and $\beta > 0$ control their regularization strengths.

To induce hardware-friendly structure in S, we reformulate (1) as follows:

$$\min_{X,L,S} \ \ell(X) + \alpha \left|L\right|_* + \underbrace{\sum_{d=1}^{D} \beta_d \Pi(S^d)}_{\Omega(S)} \quad \text{s.t. } X = L + S, \tag{2}$$

where $\Omega(S)$ denotes the structured sparsity regularizer. More specifically, we partition S into D non-overlapping groups, with S^d denoting the d-th group. Let $\Pi(\cdot)$ denote the norm applied to each group, with $\beta_d > 0$ controlling the regularization strength of group d. The partition and the norm assigned to each group jointly determine the sparsity structure promoted by $\Omega(S)$ in (2). Here, X, L, and S are jointly optimized under the task loss, allowing the model weights and both components to adapt to the prescribed sparsity structure during pretraining. The element-wise sparsity induced in SALAAD is recovered by partitioning $S \in \mathbb{R}^{n\times m}$ into $D = nm$ singleton groups, assigning the absolute-value norm to every group, and setting $\beta_d = \beta$ for all $d \in \{1, \ldots, D\}$. This instantiation gives $\Omega(S) = \beta \sum_{i=1}^{n} \sum_{j=1}^{m} |S_{ij}| = \beta |S|_1$, making the unstructured sparsity a special case of (2).

We next present three instantiations that induce sparsity patterns widely regarded as hardware-friendly. These choices are not exhaustive, and the framework can be instantiated to induce other structures as hardware capabilities and deployment requirements evolve.

Block sparsity. Assuming that $p, q \in \mathbb{Z}_+$ divide n and m respectively, we partition S into $D = {}^{nm}/_{pq}$ non-overlapping blocks $\{S^d\}_{d=1}^{D}$ of size $p \times q$. Setting $\Pi(S^d) = |S^d|_{\mathrm{F}}$ and $\beta_d = \beta$ for every block gives

$$\Omega\left(S\right) \doteq \beta \sum_{d=1}^{D} \left|S^d\right|_{\mathrm{F}}. \tag{3}$$

Each block S^d is treated as a single group, whose magnitude is measured by the Frobenius norm $|\cdot|_{\mathrm{F}}$. Penalizing the sum of these block magnitudes removes a block as a whole rather than pruning its

entries independently, thereby inducing block sparsity. This regularizer follows the group-LASSO construction of Yuan & Lin (2006), with each block treated as a group.

Row-/column-wise sparsity. Row- and column-wise sparsity are obtained from block sparsity by setting $(p, q) = (1, m)$ and $(p, q) = (n, 1)$, respectively. The resulting groups are the $D = n$ rows or $D = m$ columns of S, respectively, with $\Pi(S^d) = |S^d|_2$. Setting $\beta_d = \beta$ for every group gives

$$\Omega_{\text{row}}(S) \doteq \beta \sum_{d=1}^{D} \left|S^d\right|_2 = \beta \sum_{d=1}^{n} |S_{d,:}|_2 \,, \quad \Omega_{\text{col}}(S) \doteq \beta \sum_{d=1}^{D} \left|S^d\right|_2 = \beta \sum_{d=1}^{m} |S_{:,d}|_2 \,. \tag{4}$$

The resulting group norms induce complete row or column removal.

$N{:}M$ sparsity. Assuming that $M \in \mathbb{Z}_+$ divides m, we partition every row of S into consecutive groups of M entries, yielding $D = {}^{nm}/_{M}$ non-overlapping groups $\{S^d\}_{d=1}^{D}$. Setting $\Pi(S^d) = |S^d|_1$ gives

$$\Omega\left(S\right) \doteq \sum_{d=1}^{D} \beta_d \left|S^d\right|_1 , \tag{5}$$

where β_d specifies the regularization strength of the d-th group. The target of retaining N nonzero entries in every group of M entries defines $N{:}M$ sparsity (Zhao et al., 2024).

L-only. Fixing $S \equiv 0$ in (2) yields the L-only special case with $X = L$. We use it as a reference for tolerable block granularity across model scales; further discussion appears in Appendix B.

4 Methodology

This section describes how S-SALAAD jointly learns the model weights and their low-rank and structured sparse components during pretraining and supports compression and efficient inference for edge deployment.

4.1 Structured SLR Learning

We use ADMM (Hong et al., 2016; Candès et al., 2011; Lin et al., 2010; Ouyang et al., 2013) to jointly optimize the model weights and their structured SLR components in (2). Introducing a dual variable $Y \in \mathbb{R}^{n \times m}$ and a penalty parameter $\rho > 0$ gives the augmented Lagrangian

$$\mathcal{G}_\rho\left(X, L, S, Y\right) = \ell\left(X\right) + \alpha\left|L\right|_* + \Omega\left(S\right) + \frac{\rho}{2}\left|X - L - S + \frac{Y}{\rho}\right|_{\text{F}}^2 - \frac{1}{2\rho}\left|Y\right|_{\text{F}}^2 . \tag{6}$$

We describe the selection of ρ in Appendix D.3. Within this alternating optimization, the choice of $\Omega(S)$ determines the structured-sparsity update. At each iteration $k \geq 1$, ADMM updates the primal variables (X_k, L_k, S_k) and the dual variable Y_k as follows.

Model update. Because the model-update subproblem $\arg\min_X \mathcal{G}_\rho(X, L_{k-1}, S_{k-1}, Y_{k-1})$ does not admit a closed-form solution, we approximate its solution X_k through gradient-based optimization. Any existing optimizer can be used without modification, with the original training loss $\ell(X)$ replaced by the coupled loss

$$\ell_c(X) \doteq \ell(X) + \frac{\rho}{2}\left|X - L_{k-1} - S_{k-1} + \frac{Y_{k-1}}{\rho}\right|_{\text{F}}^2 . \tag{7}$$

Low-rank proximal update. The low-rank update is given by

$$L_k = \text{prox}_{\frac{\alpha}{\rho}|\cdot|_*}(\underbrace{X_k - S_{k-1} + {}^{Y_{k-1}}/_{\rho}}_{\overset{\text{SVD}}{=} U_k \,\text{diag}(\sigma_k) V_k^\top}) = U_k \,\text{diag}\big(\{\sigma_{k,j} - {}^{\alpha}/_{\rho}\}_+\big) V_k^\top , \tag{8}$$

where $\sigma_k = \{\sigma_{k,j}\}_j$ collects the singular values at iteration k, and $\{\cdot\}_+$ denotes the element-wise positive part.

Structured-sparsity proximal update. With $Z_k \doteq X_k - L_k + {}^{Y_{k-1}}/_{\rho}$, the sparse update can be written as the following proximal problem:

$$S_k = \arg\min_S \frac{1}{\rho}\Omega(S) + \frac{1}{2}\left|S - Z_k\right|_{\text{F}}^2 = \text{prox}_{\frac{1}{\rho}\Omega(S)}(Z_k). \tag{9}$$

The non-overlapping partition underlying $\Omega(S) = \sum_{d=1}^{D} \beta_d \Pi(S^d)$ decomposes this proximal problem into D independent group-wise subproblems. For each $d \in \{1, \dots, D\}$, let Z_k^d and S_k^d denote the components of Z_k and S_k associated with the d-th group, respectively. Under the block-sparsity regularizer and its row-/column-wise specializations, the uniform choice $\beta_d = \beta$ yields the group soft-thresholding update

$$S_k^d = \frac{1}{\rho}\Big\{\rho - \frac{\beta}{|Z_k^d|_{\mathrm{F}}}\Big\}_+ Z_k^d, \quad S_k^d \doteq 0 \text{ if } Z_k^d = 0. \tag{10}$$

Under the $N:M$ regularizer, the group-specific coefficients $\{\beta_d\}_{d=1}^{D}$ yield the element-wise soft-thresholding update within each group

$$S_k^d = \operatorname{sign}\big(Z_k^d\big) \odot \Big\{\big|Z_k^d\big| - \frac{\beta_d}{\rho}\Big\}_+, \tag{11}$$

where $\odot$ denotes element-wise multiplication.

Dual ascent. The dual variable is updated by gradient ascent on the Lagrangian $Y_k = Y_{k-1} + \rho\,(X_k - L_k - S_k)$.

4.2 End-to-End Implementation

To support large-scale training and elastic edge deployment, we address the computational cost of low-rank updates during pretraining, the zero-shot budget-aware compression of the learned components, and efficient inference on CPUs and GPUs.

Training with I-SVD. Algorithm 1 summarizes the complete training procedure, which alternates model updates with structured SLR learning. For training efficiency, we update L, S, and Y once per structured SLR learning phase. Details of the I-Controller are provided in Appendix C.1.

Algorithm 1 S-SALAAD Training

Initialization: X, L, S, Y, V, and structured sparsity $\Omega(S)$
Input: α, β, and ρ
while True **do**
  $X \leftarrow$ BACKPROPAGATION ON $\ell_c(X)$ over minibatches (7)
  $L_0 \leftarrow L, S_0 \leftarrow S, Y_0 \leftarrow Y, V_0 \leftarrow V$
  /* STRUCTURED SLR LEARNING */
  for $k = 1, \dots, K$ **do**
    $W_k \doteq X - S_{k-1} + {}^{Y_{k-1}}\!/_{\rho}$
    $U_k, \sigma_k, V_k \leftarrow$ I-SVD-STEP(W_k, V_{k-1}) Alg. 2
    $L_k \leftarrow$ L-STEP$(U_k, \sigma_k, V_k;\ \alpha, \rho)$ (8)
    $Z_k \doteq X - L_k + {}^{Y_{k-1}}\!/_{\rho}$
    $S_k \leftarrow$ S-STEP$(Z_k;\ \beta, \rho, \Omega)$ (10) or (11)
    $Y_k = Y_{k-1} + \rho(X - L_k - S_k)$
  end for
  $L \leftarrow L_K, S \leftarrow S_K, Y \leftarrow Y_K, V \leftarrow V_K$
  $\alpha, \beta \leftarrow$ I-CONTROLLER(α, β, L, S)
end while

In Algorithm 1, the low-rank update in (8) applies singular-value soft thresholding to $X_k - S_{k-1} + {}^{Y_{k-1}}\!/_{\rho}$ at iteration $k \geq 1$. Computing an exact SVD at every iteration introduces substantial training overhead as shown in Figure 1.

We therefore use I-SVD (see Algorithm 2), which performs a single step of orthogonal iteration (Golub & Van Loan, 2013) warm-started from the cached right singular basis V_{k-1}. It returns approximate triplets (U_k, σ_k, V_k) and caches V_k for the next iteration.

Algorithm 2 Iterative SVD Step

Input W, cached basis V with m rows
$(V, R) \leftarrow \mathrm{QR}\left(W^\top (WV)\right)$
$\sigma \leftarrow \sqrt{\mathbf{1}^\top ((WV) \odot (WV))}$ ▷ $\mathbf{1}$: all-ones column vector
$U \leftarrow (WV)\,\mathrm{diag}(\sigma)^{-1}$
Return (U, σ, V)

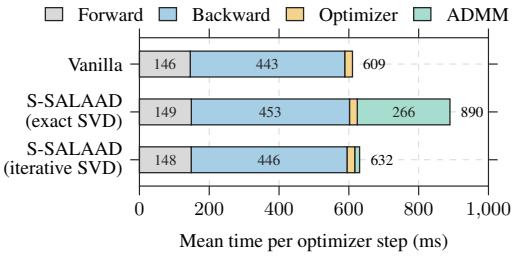


Figure 1: Per-step training-time breakdown for Llama-1B on 8 H100 GPUs.

The step applies $W_k^\top W_k$ to the cached basis V_{k-1} and orthogonalizes the product by QR, giving the new basis V_k, and the triangular factor is discarded. From the single product $W_k V_k = U_k\,\mathrm{diag}(\sigma_k)$, we obtain σ_k as the column norms and U_k as the normalized columns. Warm-starting from the previous basis enables efficient tracking across successive updates. Empirically, a single step of orthogonal iteration preserves compression quality and closely tracks the effective-rank trajectories obtained with exact SVD, as shown in Appendix C.2.

Figure 1 quantifies the resulting efficiency gain. Compared with exact SVD, iterative SVD reduces the mean time per optimizer step from 890 ms to 632 ms, a 29 % reduction. Relative to the 609 ms

vanilla baseline, this leaves only 4 % overhead. Thus, I-SVD nearly removes the ADMM overhead in S-SALAAD, making end-to-end training practical at scale.

Energy-guided alternating refinement for budget-aware compression. To enable elastic deployment, we need to compress a trained checkpoint to different parameter budgets without calibration data or further training. Given a target budget P_{tgt}, *energy-guided alternating refinement* (EGAR) first allocates low-rank and sparse capacities across projections, then refines both components within these capacities to improve reconstruction of the dense weights $\{X_i\}_{i=1}^{E}$. Both stages preserve the sparsity structure induced during pretraining. We describe block sparsity below; row-/column-wise sparsity follows by changing the partition.[2]

Stage I: Energy-guided budget allocation. To account for heterogeneity across linear projections, Stage I adaptively allocates the global parameter budget P_{tgt} across projections and between their low-rank and sparse components. The resulting retained ranks and sparse-block counts define the capacity constraints for Stage II.

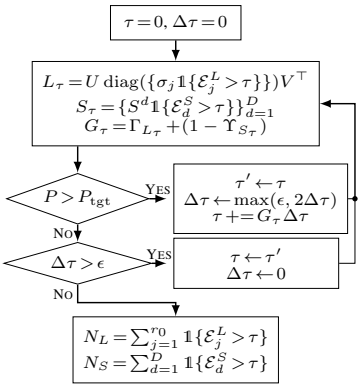


Figure 2: Pipeline of Stage I.

Allocation is guided by the energy per stored parameter. Omitting the projection index, we score singular directions and sparse blocks as

$$\mathcal{E}_j^L \doteq \sigma_j^2/_{n+m},\ j=1,\dots,r_0,\ \mathcal{E}_d^S \doteq |S^d|_{\mathrm{F}}^2/_{pq},\ d=1,\dots,D,$$

where $\{\sigma_j\}$ are the singular values of L, $r_0 = \min(n,m)$, and $\{S^d\}_{d=1}^D$ partitions S into $p \times q$ blocks. Each projection has its own threshold τ, shared by its low-rank and sparse components; candidates with scores greater than τ are retained.

For a feasible budget P_{tgt}, the coarse-to-fine search in Figure 2 adjusts the threshold τ using a minimum base step $\epsilon > 0$. Threshold increments are scaled by $G_\tau = \Gamma_{L_\tau} + (1-\Upsilon_{S_\tau})$, using the remaining-capacity statistics defined in Appendix C.1. The search returns the retained direction and block counts (N_L, N_S) for each projection, with total model parameter count $P \le P_{\mathrm{tgt}}$.

Stage II: Alternating refinement. Thresholding in Stage I can increase reconstruction error. Using the dense weight X as the reconstruction anchor, Stage II alternately refines the low-rank and sparse components L and S within the allocated capacities to reduce the reconstruction error while preserving the parameter budget. We initialize $(L,S) = (L_\tau, S_\tau)$ and define the feasible sets

$$\mathcal{R}_L \doteq \big\{A \in \mathbb{R}^{n\times m} \mid \mathrm{rank}(A) \le N_L\big\},\ \mathcal{R}_S \doteq \big\{A \in \mathbb{R}^{n\times m} \mid A \text{ has at most } N_S \text{ nonzero blocks}\big\}.$$

We then alternate the residual projections

$$L \leftarrow \mathrm{Proj}_{\mathcal{R}_L}(X-S), \quad S \leftarrow \mathrm{Proj}_{\mathcal{R}_S}(X-L). \tag{12}$$

The projections retain, respectively, the leading N_L singular components and the N_S blocks with the largest Frobenius norms of the corresponding residuals. Each update minimizes $|X-L-S|_{\mathrm{F}}^2$ over its feasible set with the other component fixed, so reconstruction error is non-increasing (see Appendix C.3). The low-rank subspace and block support remain free to change within the allocated capacities. Further details and empirical results are provided in Appendix C.3.

Dedicated CPU and GPU kernels. Among the three supported sparsity patterns, we identify block sparsity as a practical choice for elastic edge deployment, as discussed in Section 5.2. We therefore develop dedicated CPU and GPU kernels only for block-sparse inference. At inference, the compressed linear map $L+S$, with $L = AB^\top$, is applied to an input activation $h \in \mathbb{R}^m$ as

$$y = (L+S)h = (AB^\top + S)h = A(B^\top h) + Sh. \tag{13}$$

We first compute $y = AB^\top h$ using dense kernels, and then accumulate the sparse contribution Sh directly into y, avoiding a separate output buffer for the sparse path. For the sparse path, we store only the retained blocks in a contiguous layout with block-level row pointers and column indices, amortizing indexing overhead across the weights within each block. The kernels skip zero

[2]For $N:M$ sparsity, Stage I keeps the fixed-capacity S and allocates only N_L; Stage II instead uses $\mathcal{R}_S \doteq \{A \mid |A^d|_0 \le N,\ d=1,\dots,D\}$, where A^d is the d-th group of M entries.

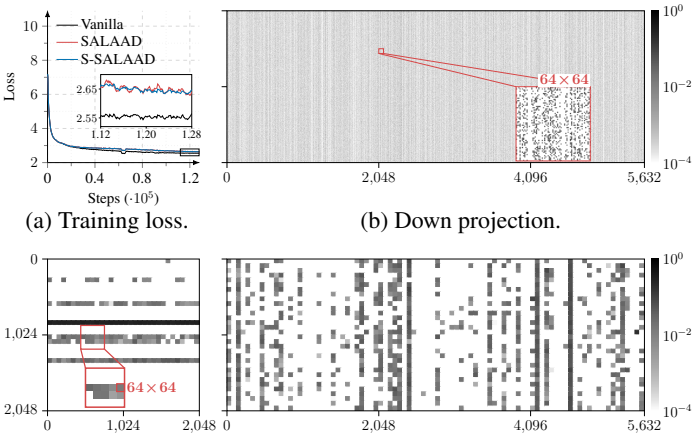
blocks and exploit contiguous access and regular computation within each retained block to reduce inference latency. We implement the GPU kernel in TileLang (Wang et al., 2025a) to perform dense tile computations on Tensor Cores, and vectorize the CPU kernel with AVX2. Kernel comparisons are reported in Appendix C.4.

5 Experiments

We pretrain Llama2 models (1B and 350M) on C4 (Raffel et al., 2020) and distill Nemotron-3-Super-120B into Qwen3-1.7B on FineWeb-Edu (Penedo et al., 2024). Dataset and model details are provided in Appendix D.2, and training settings in Appendix D.3. We compare against Wanda (Sun et al., 2024), SparseGPT (Frantar & Alistarh, 2023), SliceGPT (Ashkboos et al., 2024), SVD-LLM (Wang et al., 2025c), LORO (Mo et al., 2025), vanilla training, and SALAAD (Ma et al., 2026a). LOST (Li et al., 2025) is excluded because it requires changes to the underlying model architecture. We evaluate perplexity, zero-shot downstream accuracy, peak memory, and prefill inference latency. To assess deployment under different memory and compute constraints, we profile inference on Quadro RTX 6000 and RTX 2080 Ti GPUs and 4× AMD EPYC 7662 CPU hosts. Baseline implementations and evaluation protocols are detailed in Appendices D.4 and D.5.

5.1 Main Results

Training stability and structure induction. Figure 3a shows that S-SALAAD (64×64) on Llama-1B closely follows the loss trajectory of SALAAD, with slightly higher loss than vanilla. This modest loss penalty is accompanied by substantial memory savings and lower inference latency after compression, yielding the favorable deployment trade-offs reported in Figure 4. Stable training is also observed across sparsity patterns on Llama-350M (Figure 15a), block granularities on Llama-1B (Figure 15b), and Qwen3-1.7B distillation (Figure 15c). Figures 3b and 3d contrast element-wise and block sparsity in the down projection of the same randomly selected layer; Figure 3c shows the structured query projection. Figures 16–18 provide additional heatmaps across these settings; Figures 20–22 show effective-rank and sparsity trajectories illustrating progressive structure induction under the prescribed targets.



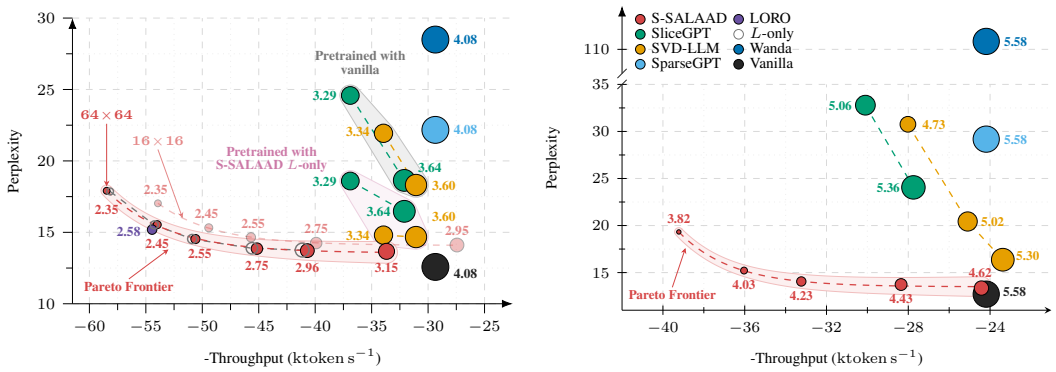
(a) Training loss. (b) Down projection.

(c) Q projection. (d) Down projection.

Figure 3: Llama-1B training and learned S in a randomly selected layer. Log shading shows entry magnitudes for SALAAD (b) and 64×64 block Frobenius norms for S-SALAAD (c,d), clipped to the displayed range.

Pareto dominance over post-training baselines. Figure 4 compares perplexity, throughput, and peak GPU memory jointly on a Quadro RTX 6000, taking the uncompressed model obtained through standard training as the vanilla-pretrained reference. Under conventional Llama pretraining, S-SALAAD establishes a superior Pareto frontier over Wanda, SparseGPT, SliceGPT, and SVD-LLM (see Figure 4a).

Indeed LORO also lies on this Pareto frontier, although each of its checkpoints is tied to a fixed model size, whereas a single S-SALAAD checkpoint supports elastic deployment across multiple parameter budgets. At the 500M operating point, block-64×64 S-SALAAD reduces peak GPU memory from 4.08 GB to 2.35 GB (42.4 %) and increases throughput from 29.35 ktoken s^{-1} to 58.47 ktoken s^{-1} (1.99×) relative to dense model inference. Plus, in compression, Wanda, SparseGPT, SliceGPT, and SVD-LLM require calibration samples to estimate statistics, whereas



(a) Llama-1B.

(b) Qwen3-1.7B.

Figure 4: Perplexity against throughput on a Quadro RTX 6000; marker radius linearly encodes peak GPU memory usage, with values annotated in GB. Lighter red denotes the 16×16 variant.

EGAR determines the compressed $L + S$ representation solely from the learned L, S, and dense weight X. The same advantage persists under cross-architecture distillation (Figure 4b): at 800M, S-SALAAD reduces peak GPU memory from 5.58 GB to 3.82 GB (31.5 %) and increases throughput from 24.19 ktoken s^{-1} to 39.22 ktoken s^{-1} (1.62×). Together, these results show that the Pareto-frontier gains extend across model families and training paradigms.

Applying SliceGPT and SVD-LLM to an S-SALAAD-pretrained checkpoint (L-only) further improves their trade-offs relative to their vanilla-pretrained counterparts, highlighting the *synergies* between S-SALAAD's learned compressibility and post-training compression methods. Additional evaluations on an RTX 2080 Ti and 4× AMD EPYC 7662 CPUs are reported in Appendix F.1.

Downstream evaluation. Table 1 compares zero-shot downstream accuracy between vanilla models and compressed variants across different parameter budgets, using the same Llama and Qwen checkpoints trained with S-SALAAD as in the Pareto-frontier evaluation (see Figure 4). Across both settings, accuracy degrades gracefully rather than collapsing as the parameter budget tightens. This trend demonstrates both that a single S-SALAAD checkpoint supports elastic deployment across capacities and that training-time induction of block-sparse SLR structure preserves the model's downstream capability.

Table 1: Zero-shot accuracy of Vanilla and S-SALAAD (in blue) across budgets.

	Llama-1B					Qwen3-1.7B				
Task	Vanilla	800M	700M	600M	500M	Vanilla	1273M	1200M	1000M	800M
PIQA	0.72	0.71	0.70	0.69	0.68	0.73	0.72	0.72	0.72	0.70
SciQ	0.79	0.76	0.76	0.76	0.77	0.92	0.92	0.91	0.90	0.89
WinoG.	0.54	0.54	0.54	0.54	0.54	0.57	0.58	0.59	0.58	0.56
HellaSwag	0.39	0.37	0.37	0.37	0.35	0.46	0.45	0.44	0.43	0.40
ARC-Easy	0.54	0.51	0.51	0.51	0.50	0.76	0.74	0.73	0.72	0.68
OBQA	0.23	0.20	0.20	0.20	0.19	0.30	0.28	0.27	0.27	0.26

5.2 Ablation Studies

Comparison of sparsity patterns. Figure 5 compares perplexity against parameter count for block, $2:4$, and row-/column-wise sparsity on Llama-350M. Although $2:4$ sparsity achieves lower perplexity at the largest budgets, its fixed sparse capacity restricts elastic compression, with perplexity rising sharply as the parameter count decreases. Block and row-/column-wise sparsity instead maintain competitive perplexity over a wider budget range.

Heatmaps of the sparse components learned with S-SALAAD reveal pronounced row- or column-aligned patterns in attention projections (Appendices E.2 and E.3). This organization helps explain why the additional pruning freedom of blocks does not yield a clear advantage over row-/column-wise sparsity in the perplexity–parameter curves. With comparable quality trade-offs, the efficient mapping of block computations onto current hardware motivates our focus on block sparsity and the development of dedicated CPU and GPU kernels. This choice reflects current hardware and deployment requirements; other patterns may become preferable as these evolve, and S-SALAAD's unified formulation accommodates such alternatives.

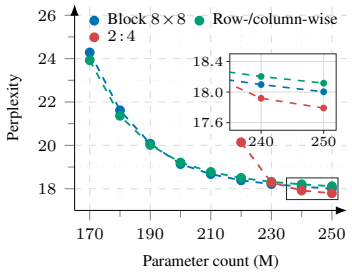
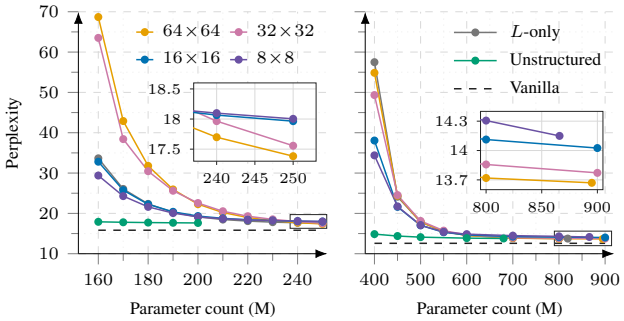


Figure 5: Perplexity vs. parameter count on Llama-350M.

Block granularity and model scale. Figure 6 compares the perplexity–parameter trade-offs of different block sizes on Llama-350M and Llama-1B, including L-only (see Appendix B). Within each model scale, the variants share the same base training configuration, with their settings for structure induction detailed in Appendix D.3. Retaining S provides representational flexibility beyond the low-rank component, helping preserve model quality as the parameter budget tightens. Finer blocks impose weaker structural constraints and allow more flexible allocation of sparse capacity during compression. This benefit is evident for the 8×8 and 16×16 variants on both Llama-350M and Llama-1B at tight budgets (Figures 6a and 6b). For example, at 500M parameters on Llama-1B, their perplexities are 17.07 and 17.03, respectively, compared with 17.86 for L-only.

This representational advantage, however, does not automatically translate into lower inference latency. Dense matrix kernels efficiently support the low-rank path, while fine-grained sparse inference introduces additional overhead. At the 500M operating point on Llama-1B, the 16×16 variant achieves lower perplexity but incurs 7.9% higher latency than L-only at approximately the same peak memory (Figures 6b and 4a). Similar overhead is observed on the RTX 2080 Ti and CPU (Figures 23a and 24a). Increasing block size trades some compression flexibility for more efficient inference. With 64×64 blocks, the perplexity–memory–latency trade-offs closely track those of L-only on both GPUs (Figures 4a and 23a), while improving on them at CPU operating points (Figure 24a). These results demonstrate that an appropriate block granularity enables the sparse component to improve deployment trade-offs beyond the L-only special case. We note that the block-size benchmark in Appendix C.4 and additional comparisons in Appendix G further characterize this trade-off. The granularity that approximately preserves the L-only perplexity–parameter trade-off increases from 16×16 on Llama-350M to 64×64 on Llama-1B (Figure 6). This trend suggests that larger models can accommodate coarser, more efficient blocks while retaining elastic compressibility, increasing the deployment benefits of structured S representations at larger scales.



(a) Llama-350M. (b) Llama-1B.

Figure 6: Perplexity versus post-compression parameter count for different block sizes on Llama.

6 Conclusion

We presented S-SALAAD, a plug-and-play framework that learns structured SLR representations during training for elastic edge deployment. Zero-shot budget-aware compression and dedicated block-sparse CPU and GPU kernels translate this learned compressibility into memory savings and efficient inference across budgets from a single checkpoint. Across Llama pretraining and Qwen distillation, S-SALAAD advances the empirical perplexity-memory-latency Pareto frontier relative to the evaluated baselines. Our analysis further links tolerable block granularity to model scale, providing empirical guidance for extending the framework to larger models.

USE OF LARGE LANGUAGE MODELS

In this work, we used generative AI tools to provide feedback on research methodology and experiments, implement methods, translate text, and clean and reformat datasets. Additionally, we used generative AI tools to create and edit plotting code and scientific figures, and to edit the manuscript for readability. We did not use generative AI tools to design the research methodology or experiments, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for or assist in writing mathematical proofs, propose or refine hypotheses, or interpret results. Synthetic-data generation and qualitative or thematic data analysis were not applicable to this work. All AI-assisted work was reviewed by the authors, and the implementations, processed data, plotting code, and resulting figures were verified by two authors against the intended methods, source data, and experimental records. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The learning objective and optimization updates are specified in Sections 3 and 4.1, and Algorithms 1 and 2 describe training and I-SVD, respectively. Section 4.2 and Appendix C document compression, inference, and supporting implementation details. Appendix D provides the datasets and tokenization, model configurations, training hyperparameters, parameter counts, baseline implementations, and evaluation protocols, including hardware, precision, input shapes, timing, and memory measurement. Model checkpoints, teacher logits, and the codebase will be released.

REFERENCES

Saleh Ashkboos, Maximilian Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. SliceGPT: Compress Large Language Models by Deleting Rows and Columns. In *International Conference on Learning Representations*, pp. 11682–11701, 2024.

Sabrina Bodmer, René Zurbrügg, Tifanny Portela, Hao Ma, Alexandre Didier, Marco Hutter, Colin Jones, and Melanie Zeilinger. Grounding Generative Policies in Physics: Optimization-Guided Diffusion for Robot Control. In *Conference on Robot Learning*, pp. 1–26, 2026.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust Principal Component Analysis? *Journal of the ACM*, 58(3):1–37, 2011.

Yueyue Deng, Pierre-Philippe J. Beaujean, Edgar An, and Edward Carlson. Task Allocation and Path Planning for Collaborative Autonomous Underwater Vehicles Operating through an Underwater Acoustic Network. *Journal of Robotics*, 2013(1):483095, 2013.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. In *Conference on Neural Information Processing Systems*, pp. 30318–30332, 2022.

European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation), 2016. Official Journal of the European Union, L 119, 1–88.

Rolf Findeisen and Frank Allgöwer. Computational Delay in Nonlinear Model Predictive Control. *IFAC Proceedings Volumes*, 37(1):427–432, 2004.

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, 2023.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The Language Model Evaluation Harness. Zenodo, version v0.4.3, 07 2024.

Gene H Golub and Charles F Van Loan. *Matrix Computations*. Johns Hopkins University Press, 2013.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge Distillation of Large Language Models. In *International Conference on Learning Representations*, pp. 32694–32717, 2024.

Tyler Gunn and John Anderson. Dynamic heterogeneous team formation for robotic urban search and rescue. *Journal of Computer and System Sciences*, 81(3):553–567, 2015.

Andi Han, Jiaxiang Li, Wei Huang, Mingyi Hong, Akiko Takeda, Pratik Jawanpuria, and Bamdev Mishra. SLTrain: a sparse plus low-rank approach for parameter and memory efficient pretraining. In *Conference on Neural Information Processing Systems*, pp. 118267–118295, 2024.

Song Han, Jeff Pool, John Tran, and William J. Dally. Learning both Weights and Connections for Efficient Neural Networks. In *Conference on Neural Information Processing Systems*, pp. 1–9, 2015.

Mingyi Hong, Zhi-Quan Luo, and Meisam Razaviyayn. Convergence Analysis of Alternating Direction Method of Multipliers for a Family of Nonconvex Problems. *SIAM Journal on Optimization*, 26(1):337–364, 2016.

Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. Language model compression with weighted low-rank factorization. In *International Conference on Learning Representations*, pp. 1–15, 2022.

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:2001.08361*, 2020.

Sampo Kuutti, Richard Bowden, Yaochu Jin, Phil Barber, and Saber Fallah. A Survey of Deep Learning Applications to Autonomous Vehicle Control. *IEEE Transactions on Intelligent Transportation Systems*, 22(2):712–733, 2021.

Kwanhee Lee, Hyeondo Jang, Dongyeop Lee, Dan Alistarh, and Namhoon Lee. The Unseen Frontier: Pushing the Limits of LLM Sparsity with Surrogate-Free ADMM. In *International Conference on Learning Representations*, pp. 29491–29512, 2026.

Jiaxi Li, Lu Yin, Li Shen, Jinjin Xu, Liwu Xu, Tianjin Huang, Wenwu Wang, Shiwei Liu, and Xilu Wang. LOST: Low-rank and Sparse Pre-training for Large Language Models. *arXiv:2508.02668*, 2025.

Zhouchen Lin, Minming Chen, and Yi Ma. The Augmented Lagrange Multiplier Method for Exact Recovery of Corrupted Low-Rank Matrices. *arXiv:1009.5055*, 2010.

Shaoshan Liu, Liangkai Liu, Jie Tang, Bo Yu, Yifan Wang, and Weisong Shi. Edge Computing for Autonomous Driving: Opportunities and Challenges. *Proceedings of the IEEE*, 107(8):1697–1716, 2019.

Hao Ma, Dieter Büchler, Bernhard Schölkopf, and Michael Muehlebach. Reinforcement learning with model-based feedforward inputs for robotic table tennis. *Autonomous Robots*, 47(8):1387–1403, 2023.

Hao Ma, Sabrina Bodmer, Andrea Carron, Melanie Zeilinger, and Michael Muehlebach. Constraint-Aware Diffusion Guidance for Robotics: Real-Time Obstacle Avoidance for Autonomous Racing. In *Conference on Robot Learning*, pp. 1–19, 2025.

Hao Ma, Melis Ilayda Bal, Liang Zhang, Bingcong Li, Niao He, Melanie Zeilinger, and Michael Muehlebach. SALAAD: Sparse And Low-Rank Adaptation via ADMM for Large Language Model Inference. In *International Conference on Machine Learning*, pp. 1–21, 2026a.

Hao Ma, Melanie Zeilinger, and Michael Muehlebach. Stochastic Online Optimization for Cyber-Physical and Robotic Systems. *Machine Learning*, 115(1):11, 2026b.

Mehdi Makni, Xiang Meng, and Rahul Mazumder. 3BASiL: An Algorithmic Framework for Sparse plus Low-Rank Compression of LLMs. In *Conference on Neural Information Processing Systems*, pp. 173917–173947, 2025.

Asit Mishra, Jorge Albericio Latorre, Jeff Pool, Darko Stosic, Dusan Stosic, Ganesh Venkatesh, Chong Yu, and Paulius Micikevicius. Accelerating Sparse Deep Neural Networks. *arXiv:2104.08378*, 2021.

Zhanfeng Mo, Long-Kai Huang, and Sinno Jialin Pan. Parameter and Memory Efficient Pretraining via Low-rank Riemannian Optimization. In *International Conference on Learning Representations*, pp. 7907–7936, 2025.

Fang Nan, Hao Ma, Qinghua Guan, Josie Hughes, Michael Muehlebach, and Marco Hutter. Efficient Model-Based Reinforcement Learning for Robot Control via Online Optimization. *The International Journal of Robotics Research*, 0(0):1–18, 2026.

Patrik Okanovic, Sameer Deshmukh, Grzegorz Kwasniewski, Yi Zhu, Haruto Fujii, Sakina Fatima, Maciej Besta, Kentaro Katayama, Takumi Honda, Yusuke Nagasaka, et al. BLaST: High Performance Inference and Pretraining using BLock Sparse Transformers. *arXiv:2507.03117*, 2025.

Hua Ouyang, Niao He, Long Tran, and Alexander Gray. Stochastic Alternating Direction Method of Multipliers. In *International Conference on Machine Learning*, pp. 80–88, 2013.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale. In *Conference on Neural Information Processing Systems*, pp. 30811–30849, 2024.

Hao Peng, Xin Lv, Yushi Bai, Zijun Yao, Jiajie Zhang, Lei Hou, and Juanzi Li. Pre-training Distillation for Large Language Models: A Design Space Exploration. In *Annual Meeting of the Association for Computational Linguistics*, pp. 3603–3618, 2025.

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.

Weisong Shi, Jie Cao, Quan Zhang, Youhuizi Li, and Lanyu Xu. Edge Computing: Vision and Challenges. *IEEE Internet of Things Journal*, 3(5):637–646, 2016.

Kang G. Shin and Xianzhong Cui. Computing Time Delay and Its Effects on Real-Time Control Systems. *IEEE Transactions on Control Systems Technology*, 3(2):218–224, 1995.

Milica Stojanovic and James C. Preisig. Underwater Acoustic Communication Channels: Propagation Models and Statistical Characterization. *IEEE Communications Magazine*, 47(1):84–89, 2009.

Mingjie Sun, Zhuang Liu, Anna Bair, and Zico Kolter. A Simple and Effective Pruning Approach for Large Language Models. In *International Conference on Learning Representations*, pp. 4942–4964, 2024.

Lei Wang, Yu Cheng, Yining Shi, Zhengju Tang, Zhiwen Mo, Wenhao Xie, Lingxiao Ma, Yuqing Xia, Jilong Xue, Fan Yang, et al. TileLang: A Composable Tiled Programming Model for AI Systems. *arXiv:2504.17577*, 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 *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, 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*, pp. 19299–19319, 2025c.

Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning Structured Sparsity in Deep Neural Networks. In *Conference on Neural Information Processing Systems*, pp. 1–9, 2016.

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:2505.09388*, 2025.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.

Kang Zhao, Tao Yuan, Han Bao, Zhenfeng Su, Chang Gao, Zhaofeng Sun, Zichen Liang, Liping Jing, and Jianfei Chen. Beyond 2:4: Exploring V:N:M Sparsity for Efficient Transformer Inference on GPUs. *arXiv:2410.16135*, 2024.

A Motivation for Structured Sparsity During Pretraining

A.1 Memory Savings Do Not Guarantee Lower Inference Latency

We compare dense inference with the unstructured SLR representation used by SALAAD on a synthetic Llama-1B MLP down projection of size 2048×5632. The benchmark uses a 15 % rank ratio and 5 % nonzero density, with the sparse component stored as a PyTorch `torch.sparse_csr_tensor`. Both representations are profiled on a Quadro RTX 6000 with batch size 128 and sequence length 256. As shown in Figure 7, static weight memory falls from 23.07 MB to 10.49 MB, a 54.5 % reduction, while throughput falls to 2.68 % of the dense reference. For this fixed workload, the throughput ratio implies a projection inference latency 37.3 times that of the dense baseline. Unstructured sparse components incur indexing and irregular memory-access overheads that can offset the computational savings from sparsity (Mishra et al., 2021). This example illustrates why reduced weight storage alone does not ensure efficient inference.

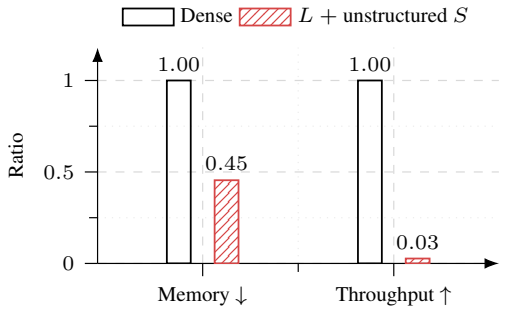


Figure 7: Static weight memory and throughput of dense inference and the unstructured SLR representation used by SALAAD. Both metrics are normalized to their respective dense baselines. The benchmark uses a synthetic Llama-1B MLP down projection with a 15 % rank ratio and 5 % nonzero density.

A.2 Quality Degradation from Imposing Block Structure After Pretraining

We test post-hoc block compression by applying energy-guided alternating refinement (EGAR; Section 4.2) to the same SALAAD Llama-1B checkpoint pretrained with unstructured sparsity, under different block constraints and parameter budgets. Table 2 reveals two trends. First, at a fixed budget, perplexity rises as the imposed blocks become coarser. Second, with coarser blocks, perplexity increases more steeply as the parameter budget decreases.

These trends are consistent with the reduced pruning flexibility of coarser blocks. Grouping more weights into a single pruning unit restricts which weights can be retained at a given budget, limiting the compressed model's representational capacity. A checkpoint pretrained with unstructured sparsity has not adapted to these constraints, making post-hoc pruning more damaging as the budget tightens. These results motivate inducing the target structure during pretraining, allowing the sparse and low-rank components to adapt to it under the task loss.

Table 2: Post-compression perplexity of the same unstructured SALAAD Llama-1B checkpoint when block structure of size $p \times p$ is imposed during compression, at three parameter budgets P_{tgt}. The first row is the element-wise reference.

Block size	$P_{\text{tgt}} = 500$M	$P_{\text{tgt}} = 600$M	$P_{\text{tgt}} = 700$M
1×1	14.11	13.86	13.78
2×2	17.29	14.93	14.19
4×4	87.44	28.89	18.28
8×8	456.52	116.33	43.58
16×16	1497.68	351.39	98.83

B L-ONLY S-SALAAD

The L-only variant is the $S \equiv 0$ special case of S-SALAAD. It retains ADMM's coupled updates of X and L under the constraint $X = L$, with the proximal update (8) explicitly removing small singular values from L. This structure-learning procedure goes beyond directly adding a low-rank regularizer to the task loss.

Figure 8 compares this variant with direct subgradient optimization of $\ell(X) + \lambda|X|_*$. With $\lambda = 10^{-6}$, perplexity is comparable to that of L-only S-SALAAD at the largest budget but rises sharply as the budget decreases. This behavior is consistent with low-rank structure that is too weak to support aggressive compression. Increasing λ to 10^{-5} strengthens rank regularization and yields nearly constant perplexity across the tested budgets, supporting elastic compression at the cost of higher perplexity throughout. L-only S-SALAAD maintains low perplexity over the same budget range, providing a better trade-off between model quality and elastic compressibility than these direct regularization baselines.

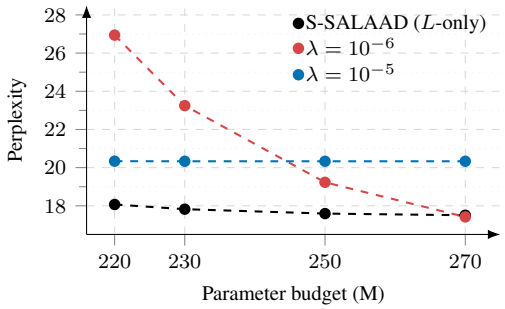


Figure 8: Post-compression perplexity on Llama-350M.

C ADDITIONAL TRAINING AND COMPRESSION DETAILS

C.1 ADAPTIVE I-CONTROLLER

S-SALAAD follows the integral controller (I-controller) update rule of SALAAD (Ma et al., 2026a), which we briefly restate here for completeness. We first define the effective-rank ratio and group sparsity used by the controller.

For a low-rank component $L \in \mathbb{R}^{n\times m}$, let $r_0 \doteq \min(n, m)$ and let $\sigma_1 \geq \cdots \geq \sigma_{r_0} \geq 0$ denote its singular values. For $L \neq 0$ and an energy-coverage level $\gamma \in (0, 1]$, we define

$$i_\gamma \doteq \min\left\{ i \in \{1, \ldots, r_0\} : \frac{\sum_{j=1}^{i} \sigma_j^2}{\sum_{j=1}^{r_0} \sigma_j^2} \geq \gamma \right\}, \quad \Gamma_L \doteq \frac{i_\gamma}{r_0}. \tag{14}$$

When $L = 0$, we set $i_\gamma = \Gamma_L = 0$. Thus, Γ_L is the fraction of available singular directions required to retain a fraction γ of the energy in L. For a sparse component S partitioned into the D non-overlapping groups $\{S^d\}_{d=1}^{D}$ of Ω, its sparsity is the fraction of groups that are zero,

$$\Upsilon_S \doteq \frac{\sum_{d=1}^{D} \mathbb{1}\{|S^d|_{\mathrm{F}} = 0\}}{D}. \tag{15}$$

Evaluating the statistic over groups rather than entries matches the granularity at which the S-update removes parameters; for $D = nm$ singleton groups, this reduces to the element-wise sparsity $1 - |S|_0/nm$. Both statistics are used by the I-controller during training and by Stage I of EGAR, which recomputes them after every thresholding step and uses $G \doteq \Gamma_L + (1 - \Upsilon_S)$ to adapt the next threshold update to the current low-rank and sparse components.

After each ADMM round, the controller adjusts the regularization strengths α and β so that the induced effective-rank ratio Γ_L and group sparsity Υ_S track targets $(\widehat{\Gamma}, \widehat{\Upsilon})$. Given controller step

sizes $(\Delta_\alpha, \Delta_\beta)$, the update is

$$\alpha \leftarrow \alpha + \rho \left(\Gamma_L - \widehat{\Gamma}\right) \Delta_\alpha, \quad \beta \leftarrow \beta + \rho \left(\widehat{\Upsilon} - \Upsilon_S\right) \Delta_\beta. \tag{16}$$

When Γ_L exceeds its target, increasing α raises the low-rank proximal threshold α/ρ; when Υ_S falls below its target, increasing β raises β/ρ and induces more zero groups. The corresponding penalty decreases in the opposite case. The factor ρ makes the increments of the effective thresholds independent of the ADMM penalty, since dividing (16) by ρ yields the increments $\left(\Gamma_L - \widehat{\Gamma}\right) \Delta_\alpha$ and $\left(\widehat{\Upsilon} - \Upsilon_S\right) \Delta_\beta$. Under the $N:M$ regularizer, where Ω carries group-specific coefficients $\{\beta_d\}_{d=1}^{D}$, the β-update in (16) is applied per group, with Υ_S replaced by the element-wise sparsity $\Upsilon_{S^d} = 1 - |S^d|_0/M$ within the d-th group and target $\widehat{\Upsilon} = 1 - N/M$; the α-update is unchanged. The selection of ρ and the controller step sizes is described in Appendix D.3.

C.2 Fidelity of Iterative SVD

Figures 9 and 10 compare I-SVD with exact SVD beyond the training-time results reported in the main text. I-SVD preserves both the post-compression validation perplexity and the effective-rank trajectories of exact SVD, confirming that the approximation does not degrade the learned decomposition.

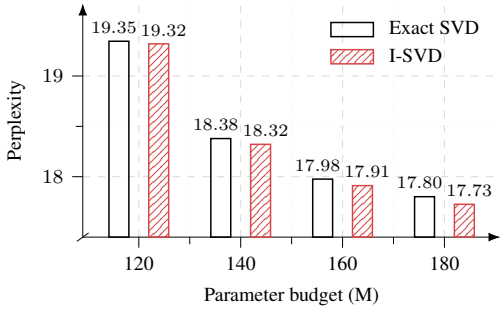
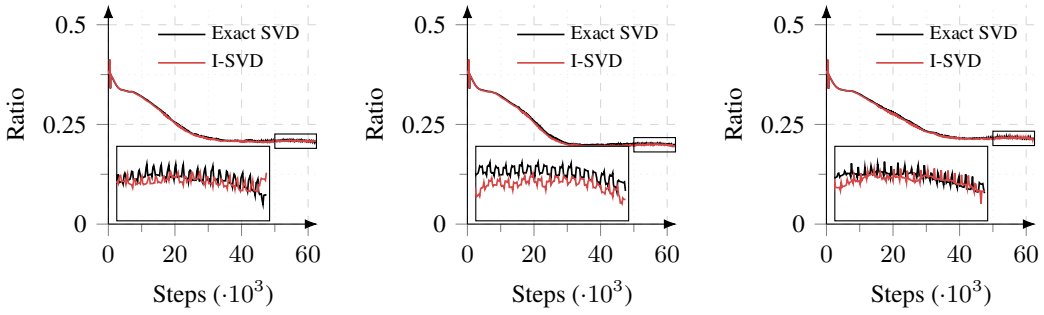


Figure 9: Post-compression perplexity for exact and iterative SVD on Llama-350M at matched budgets.



(a) Layer 4. (b) Layer 8. (c) Layer 14.

Figure 10: Γ_L trajectories for exact and iterative SVD on the down projections of three representative Llama-350M layers. Each inset magnifies the converged regime, where the two curves are otherwise indistinguishable.

C.3 Compression by Energy-Guided Alternating Refinement

The following EGAR ablations use Llama-1B pretrained with unstructured SALAAD (Ma et al., 2026a). Figure 11 traces Stage II on Llama-1B at a 500M parameter budget. Starting from the

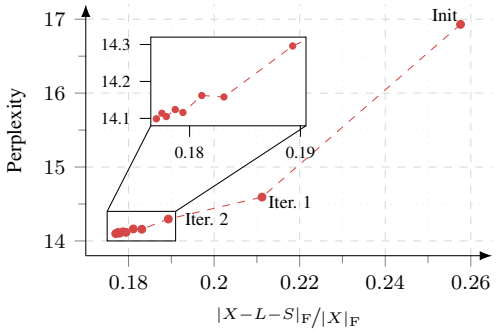


Figure 11: Alternating refinement trajectory on Llama-1B pretrained with unstructured SALAAD at a 500M parameter budget.

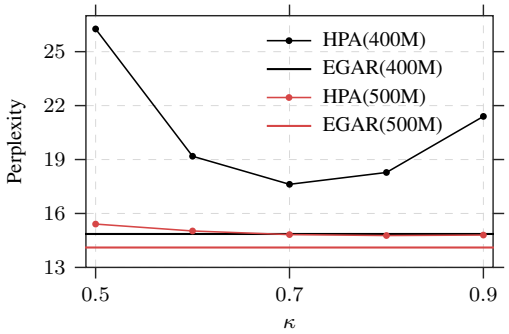
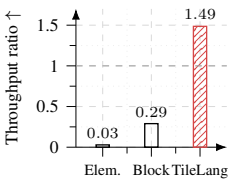


Figure 12: HPA versus EGAR on Llama-1B pretrained with unstructured SALAAD at 400M and 500M parameter budgets.

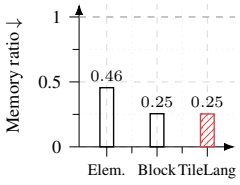
allocation produced by Stage I, alternating refinement monotonically reduces the normalized reconstruction error, while validation perplexity falls sharply and then stabilizes with minor fluctuations. Most of the improvement occurs within the first iterations, and the procedure converges in fewer than 10 iterations, taking only a few minutes in total. We compare EGAR with *homomorphic parameter allocation* (HPA), the greedy pruning strategy used in SALAAD (Ma et al., 2026a). HPA assumes homomorphic shrinkage across linear projections and performs single-pass greedy pruning after splitting P_{tgt} between sparse and low-rank components with a rigid ratio κ, thereby ignoring the heterogeneity across linear projections and potentially degrading performance. Both methods compress the same Llama-1B checkpoint pretrained with unstructured SALAAD, as shown in Figure 12. At the 400M and 500M parameter budgets, EGAR achieves lower perplexity than HPA for all evaluated values of κ between 0.5 and 0.9.

C.4 Kernel Micro-Benchmark

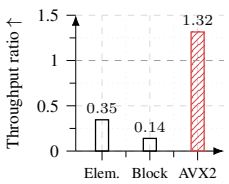
On a randomly sparsified Llama-1B linear projection (Figure 13), unstructured sparsity incurs prohibitive compute overhead and moderate memory overhead. Stock PyTorch block-sparse SpMM[3] is in beta version. In practice we find it uncompetitive with the dense baseline on either platform. Our custom TileLang (GPU) and AVX2 (CPU) kernels realise the structural savings and deliver measured speedups.



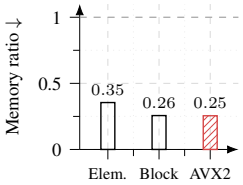
(a) GPU throughput.



(b) GPU static weight memory.



(c) CPU throughput.



(d) CPU static weight memory.

Figure 13: Kernel efficiency on a randomly sparsified Llama-1B MLP down projection at 15 % rank ratio and 5 % density, normalized to dense execution. Elem. and Block denote PyTorch's `torch.sparse_csr_tensor` and `torch.sparse_bsr_tensor` backends, respectively. In the block-sparse setting both backends run at 16×16 on the GPU and 8×8 on the CPU, the smallest block size each platform supports.

Block size and inference speed. Figure 14 compares 16×16, 32×32, and 64×64 blocks on randomly generated SLR projections with rank 307 and a target nonzero density of 5 %; block masks are sampled independently at each size. With our TileLang kernel, throughput relative to

[3]PyTorch 2.13.0 with the CUDA 12.9 build.

dense increases from $1.49\times$ to $1.87\times$ and $2.33\times$, respectively, giving 64×64 blocks $1.57\times$ the throughput of 16×16 blocks in this setting. PyTorch BSR remains slower than dense at all three sizes, highlighting the role of the kernel in realizing the benefits of coarser blocks.

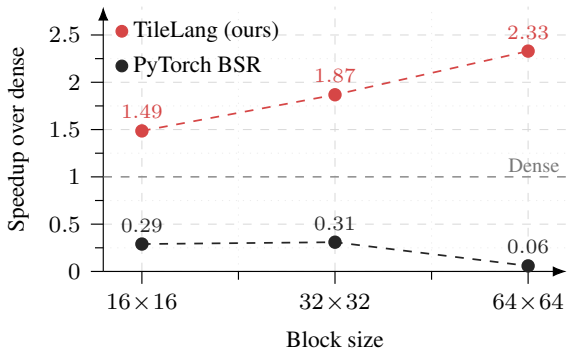


Figure 14: Block-size sweep on random SLR weights for a 2048×5632 Llama-1B down projection, measured on a Quadro RTX 6000 in FP16 with batch size 128 and sequence length 256. Throughput is normalized to dense.

D Implementation Details

D.1 Relevance to Edge Deployment

This work focuses on reducing the memory footprint and inference latency of large models for edge deployment. Our evaluation uses single-GPU inference on a Quadro RTX 6000 with 24 GB of GPU memory and an RTX 2080 Ti with 11 GB of GPU memory, together with CPU-only inference on $4\times$ AMD EPYC 7662 CPUs with 32 GB of RAM. Edge computing encompasses computation near data sources, including local compute nodes (Shi et al., 2016). We use these workstation/server platforms as proxies for edge inference in terms of memory capacity and available acceleration, the resource constraints targeted by S-SALAAD. Energy consumption and thermal management are outside the scope of this study.

D.2 Datasets and Model Configurations

Datasets and tokenization. For Llama 2 pretraining from scratch we use the C4 corpus (Raffel et al., 2020) tokenised with `t5-base`. For the cross-architecture distillation of Qwen3-1.7B we use FineWeb-Edu (Penedo et al., 2024), tokenised with the teacher's Nemotron tokenizer so that student and teacher share a token space.

Llama 2 pretraining sizes. Table 3 lists the Llama 2 architectures we pretrain. Llama-1B supports the main block-sparse experiments, while Llama-350M is used for comparisons among sparsity patterns. We compare block granularities across both model sizes.

Table 3: Model architectures used for Llama 2 pretraining and Qwen3 distillation.

Setting	**Llama-350M**	**Llama-1B**	**Qwen3-1.7B**
Hidden size	1024	2048	2048
Intermediate (FFN) size	2816	5632	6144
Layers	24	24	28
Attention heads (Q / KV)	16 / 16 (MHA)	32 / 32 (MHA)	16 / 8 (GQA)
Head dimension	64	64	128
Vocabulary size	32,128	32,128	131,072
Activation	SwiGLU		
Normalisation	RMSNorm ($\varepsilon = 10^{-6}$), pre-norm		
QK-normalisation	no	no	yes
Positional encoding	RoPE, $\theta = 10^4$		RoPE, $\theta = 10^6$
Tied I/O embeddings	no	no	yes

Qwen3-1.7B distillation student. We adopt the standard Qwen3-1.7B backbone (Yang et al., 2025) with a single modification (see Table 3): the input/output embedding vocabulary is swapped from Qwen3's native one to the teacher's Nemotron tokenizer, so that student and teacher share a token space and the top-K KL distillation loss is well-defined without a re-tokenisation step.

D.3 Training Configurations

Pretraining from scratch. Table 4 summarizes the shared and model-specific settings for Llama 2 pretraining.

Cross-architecture distillation. The student uses the teacher's tokenizer and is trained with the distillation loss (Peng et al., 2025)

$$\mathcal{L}_{\text{distill}} = (1 - \alpha_{\text{CE}})\ \text{KL}_{\text{top-}K}(p_{\text{T}} \,\|\, p_{\text{S}}) + \alpha_{\text{CE}}\ \text{CE}(y, p_{\text{S}}),$$

where p_{T} and p_{S} denote the teacher and student next-token distributions, respectively, and y is the ground-truth next token. Table 5 lists the settings used for the cross-architecture distillation.

Table 4: Llama 2 pretraining settings.

Setting	350M	1B
Optimizer	Adam	
(β_1, β_2)	$(0.9, 0.95)$	
ε	10^{-8}	
LR schedule	cosine, 10 % warmup, min-LR ratio 0.1	
Batch size	512 sequences	
Sequence length	256	
Tokenizer	`t5-base`	
Precision	bf16 mixed	
Peak LR	10^{-3}	5×10^{-4}
Total steps	62,250	128,000
Tokens (B)	8.16	16.78

Table 5: Cross-architecture distillation.

Setting	Value
Teacher	Nemotron-3-Super-120B (NVFP4)
Corpus	FineWeb-Edu
Tokenizer	Nemotron (teacher's)
α_{CE}	0.1
Top-K	10
Optimizer	AdamW
(β_1, β_2)	$(0.9, 0.95)$
Peak LR	4×10^{-4}
Schedule	cosine, 10 % warmup, min-LR ratio 0.1
Total steps	180,000
Batch size	512 sequences
Sequence length	512

Hyperparameter settings. We use a target effective-rank ratio $\widehat{\Gamma} = 0.15$ and target sparsity $\widehat{\Upsilon} = 0.95$. The L-only special case retains the same rank target but omits the sparse target and β-update, while $2:4$ sparsity uses $\widehat{\Gamma} = 0.03$ and $\widehat{\Upsilon} = 0.5$. We fix the energy-coverage level $\gamma = 0.999$ throughout. For the row-/column-wise variant, we sparsify rows in the query, key, value, gate, and up projections and columns in the output and down projections.

We set ρ according to the empirical scaling rule proposed in SALAAD (Ma et al., 2026a),

$$\rho \propto \frac{1}{E\sqrt{nm}}, \tag{17}$$

where E is the number of selected linear projections and $n \times m$ is the shape of the current projection's weight matrix. The unstructured baseline uses the Δ_β setting from SALAAD. To induce structured sparsity, we increase Δ_β and correspondingly decrease Δ_α. Within each sparsity pattern, we make only minor adjustments to the controller step sizes Δ_α and Δ_β, whose values are reported in Table 6.

Our experiments indicate that, within a given sparsity variant, Δ_α and Δ_β can generally be reused across model scales, with ρ alone adjusted according to the scaling rule above. This observation is consistent with SALAAD's hyperparameter-scaling strategy (Ma et al., 2026a) and provides an empirical basis for extending S-SALAAD to larger models.

Table 6: Controller step sizes across model scales and sparsity variants.

Variant	**Llama-350M**		**Llama-1B**		**Qwen3-1.7B**	
	Δ_α	Δ_β	Δ_α	Δ_β	Δ_α	Δ_β
Unstructured SALAAD	0.20	0.001	0.20	0.001	—	—
Block-8×8	0.04	0.01	0.045	0.01	—	—
Block-16×16	0.04	0.01	0.04	0.01	—	—
Block-32×32	0.04	0.01	0.04	0.01	—	—
Block-64×64	0.02	0.02	0.02	0.02	0.02	0.02
L-only ($S \equiv 0$)	0.03	—	0.03	—	—	—
$2:4$ semi-structured	0.10	0.025	—	—	—	—
Row-/column-wise	0.04	0.04	—	—	—	—

D.4 Evaluation and Baselines

Quality evaluation. Except for the separately sourced LORO reference described below, perplexity is measured on held-out tokens of the training corpus—C4 for Llama and FineWeb-Edu for Qwen3-1.7B—under the corresponding evaluation protocol. Perplexities are compared within each model family, not across the two corpora.

We evaluate zero-shot downstream accuracy on ARC-Easy, OpenBookQA, WinoGrande, HellaSwag, PIQA, and SciQ using the LM Evaluation Harness (Gao et al., 2024).

Baseline implementations and profiling. Vanilla training and SALAAD (Ma et al., 2026a) provide training references, while the L-only special case ($S \equiv 0$) provides a reference for block granularity. EGAR produces variants at multiple parameter budgets from each S-SALAAD checkpoint without calibration data. All compressed models are evaluated without additional fine-tuning for fair comparison.

For Wanda and SparseGPT, we use the measured latency and throughput of the corresponding dense models, as the evaluated hardware lacks native acceleration for $2:4$ sparsity. This illustrates that the latency benefits of a sparsity pattern are constrained by support from the target hardware.

We use the SVD-LLM V2 implementation (Wang et al., 2025b) for the SVD-LLM baseline. For LORO, we use its published Llama-1B perplexity of 15.19 at rank 512 after training on 13.1B C4 tokens (Mo et al., 2025). We profile its deployment form locally by replacing the seven projections in each layer of our Llama-1B backbone with pairs of rank-512 linear factors, retaining dense embeddings and the output head. This backbone uses FFN width 5632, compared with 5461 in the published LORO model, so the combined quality–cost point serves as an external reference.

SliceGPT and SVD-LLM inference costs are likewise profiled with synthetic models of the corresponding compressed dimensions; for Llama-1B, the same cost measurements are used for their standard-pretrained and S-SALAAD-pretrained variants. The synthetic models use the same precision, input shapes, and forward-timing path as the checkpoint-based models, with the corresponding dense or factorized projection operators. For RTX 2080 Ti profiling of L-only and the 16×16 and 32×32 variants, we retain the compressed components and randomly initialize the remaining parameters, while evaluating perplexity on the corresponding trained checkpoints.

D.5 Profiling Environment and Protocol

The main GPU results use a Quadro RTX 6000, with additional results on an RTX 2080 Ti. Table 7 details the environment for the Quadro RTX 6000 and CPU measurements; CPU runs use 4× AMD EPYC 7662 CPUs.

Table 7: Profiling environment for the Quadro RTX 6000 and CPU measurements.

Component	Value
GPU	NVIDIA Quadro RTX 6000 (Turing, TU102)
GPU memory	24 GB GDDR6
GPU driver / CUDA	NVIDIA driver 550.x / CUDA 12.9
CPU	4× AMD EPYC 7662 CPUs
RAM	32 GB
PyTorch	2.13.0 (cu129 build)
TileLang	0.1.12

Inference profiling protocol. Table 8 specifies the fixed input shapes and repetition counts used for the Pareto-frontier measurements. Each run uses 10 untimed warmup passes followed by 20 timed passes, except for Qwen on CPU, which uses 3 warmup passes and 6 timed passes. GPU profiling uses FP16 on both GPU models, while CPU profiling uses FP32 on the 4× AMD EPYC 7662 CPUs. Models run in evaluation mode with gradients and KV caching disabled. GPU profiling enables `torch.compile` in `reduce-overhead` mode; CPU profiling uses eager inference.

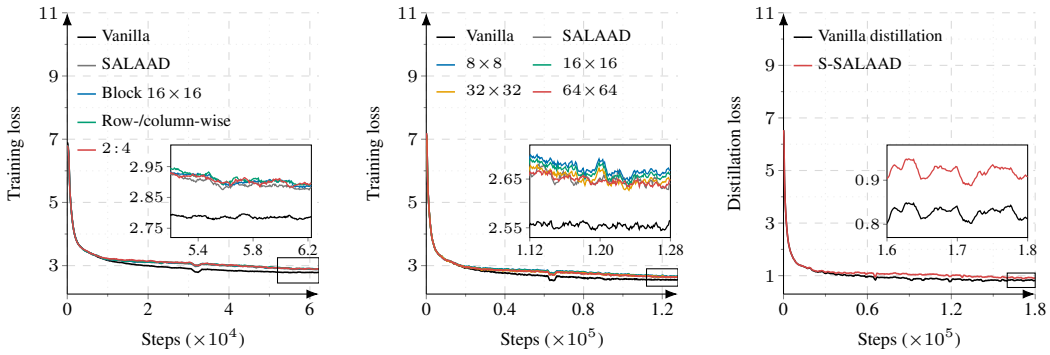
After untimed warmup, we measure each forward pass with a wall-clock timer, synchronizing CUDA immediately before and after the timed region on GPUs, and report the median latency. Throughput is the number of input tokens per batch divided by this median latency. We report results only after latency and throughput have stabilized; repeated profiling runs show negligible variation. Data loading, input transfer, compilation, and loss evaluation are excluded from timing.

The timed forward computes logits only for the final two positions of each sequence; the perplexities plotted in the figures come from separate held-out evaluations, rather than the auxiliary last-token estimates recorded during profiling. These measurements characterize prefill computation and do not include autoregressive decoding or complete request latency.

Table 8: Input shapes and repetition counts for inference profiling.

Hardware	Model	Batch × length	Warmup	Timed
Quadro RTX 6000	Llama-1B	128×256	10	20
	Qwen3-1.7B	32×512	10	20
RTX 2080 Ti	Llama-1B	128×256	10	20
	Qwen3-1.7B	32×512	10	20
4× AMD EPYC 7662 CPUs	Llama-1B	8×64	10	20
	Qwen3-1.7B	8×64	3	6

Peak-memory measurement. Peak memory is measured separately with an eager prefill pass that retains logits only at the final sequence position, without evaluating the loss. On GPUs, we reset the peak-memory counter and report peak GPU memory usage as the maximum allocated tensor memory, including resident model weights, buffers, and sparse indices; allocator-reserved memory is excluded. On CPU, we report peak RAM usage as the stored parameter and buffer sizes plus the increase in the process's peak resident memory during the forward pass. This measures model storage and forward-pass working memory without compilation allocations or a persistent KV cache.



(a) Llama-350M. (b) Llama-1B. (c) Qwen3-1.7B.

Figure 15: Task-loss trajectories for (a) different sparsity patterns on Llama-350M, (b) different block sizes on Llama-1B, and (c) Qwen3-1.7B distillation. Curves use a moving average over approximately 1,000 training steps and end at the last available logged step.

E Training Dynamics and Learned Sparse Structure

E.1 Training Loss

Figure 15 compares loss trajectories across model sizes, sparsity patterns, block granularities, and training settings. On Llama-350M, we examine block-16×16, row-/column-wise, and $2:4$ sparsity under the settings in Appendix D.3, with vanilla pretraining and unstructured SALAAD as references (Figure 15a). For Llama-1B, Figure 15b compares block sizes 8×8, 16×16, 32×32, and 64×64 with the same training references. The curves come from the runs corresponding to the released checkpoints, retaining their original training configurations and controller gains. Figure 15c compares vanilla distillation with block-64×64 S-SALAAD for the Qwen3-1.7B student over 180,000 steps, under the settings in Appendix D.3. Both runs use the same distillation objective, data, optimizer, and training schedule. The displayed task loss excludes the structural coupling penalty. These results show that introducing structured sparsity preserves stable convergence across the evaluated settings, despite a modest increase in task loss relative to vanilla training. In the Llama pretraining experiments, S-SALAAD also maintains loss trajectories comparable to those of unstructured SALAAD.

E.2 Learned Sparsity Patterns

All panels in Figures 16, 17, and 18 preserve matrix aspect ratios and share a logarithmic scale for element-wise $|S|$, clipped to $[10^{-4}, 1]$, with zeros shown in white.

Figure 16 displays the learned sparse components of randomly selected projections in Llama-350M for the three training variants in Figure 15a. For the $2:4$ training variant, the displayed components have approximately 50% density, while occupancy varies across individual four-entry groups. The regularizer and I-Controller encourage the $N:M$ target during training, while projection in EGAR Stage II enforces at most N nonzero entries per group of M. Figure 17 displays the learned sparse components of randomly selected projections in Llama-1B for the four block granularities in Figure 15b. The insets show 6×6 native blocks around a common matrix location, making the local block structure visible at each granularity. Figure 18 shows that S-SALAAD induces the intended 64×64 block structure in randomly selected projections of Qwen3-1.7B during distillation.

The block-sparse components in Figure 16 also exhibit row- or column-aligned clustering resembling that of the row-/column-wise variant. This organization persists across block granularities (Figure 17) and is also observed in Qwen3-1.7B under a different model architecture and training setting (Figure 18). We conjecture that this recurring organization reflects an intrinsic sparsity structure associated with semantic representations in Transformers. A detailed investigation of this hypothesis is beyond the scope of this work.

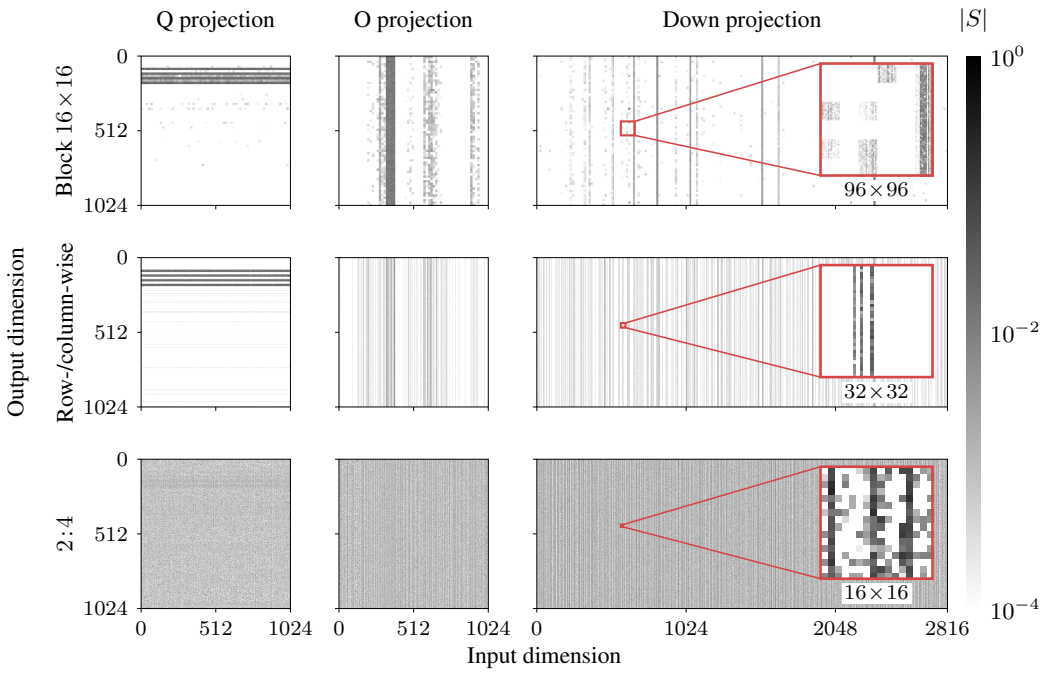


Figure 16: Learned sparse components S in layer 16 of Llama-350M. Rows show block-16×16, row-/column-wise, and $2:4$ training; columns show the Q, O, and down projections. Insets enlarge the red-boxed regions of each down projection on the same color scale; their labels give the crop dimensions in weights. The displayed matrices are taken directly from the trained checkpoints; the $2:4$ row shows the learned components before any additional structural projection.

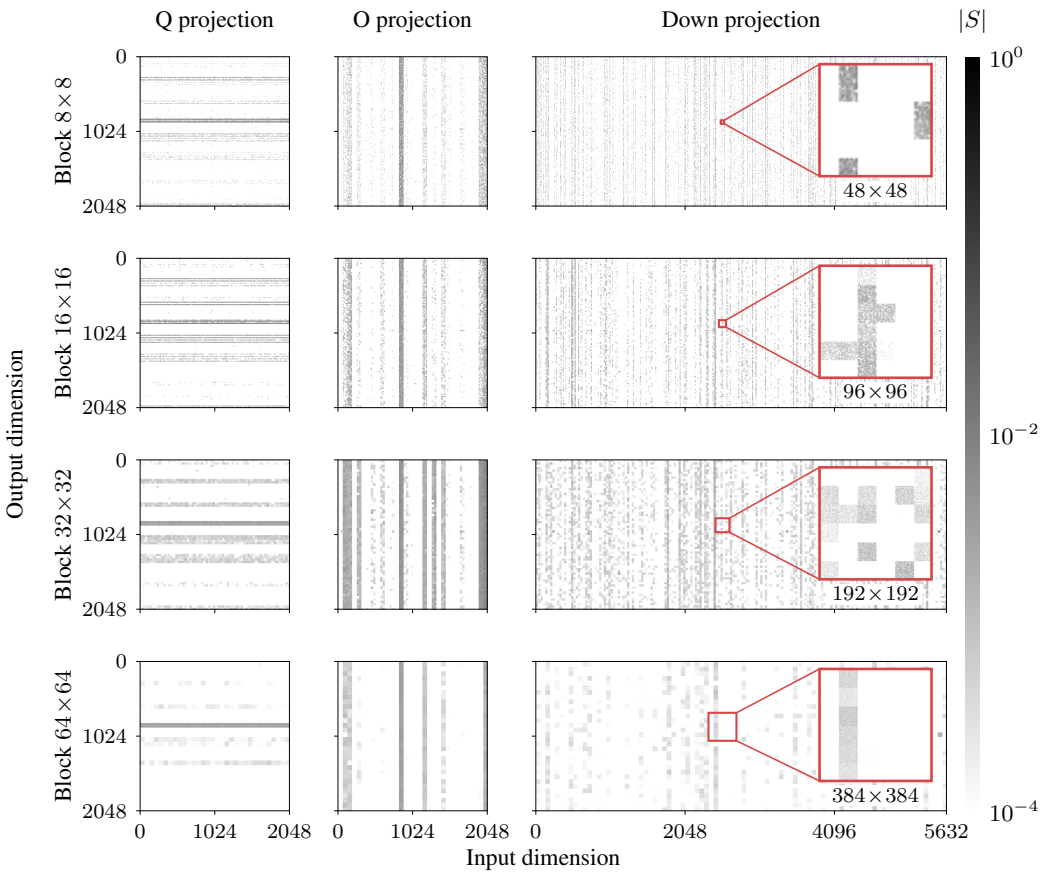


Figure 17: Learned sparse components S in layer 16 of Llama-1B across four block sizes. Columns show the Q, O, and down projections. Insets enlarge the red-boxed down-projection regions on the same color scale; labels give the crop dimensions in weights. Each inset contains 6×6 native blocks.

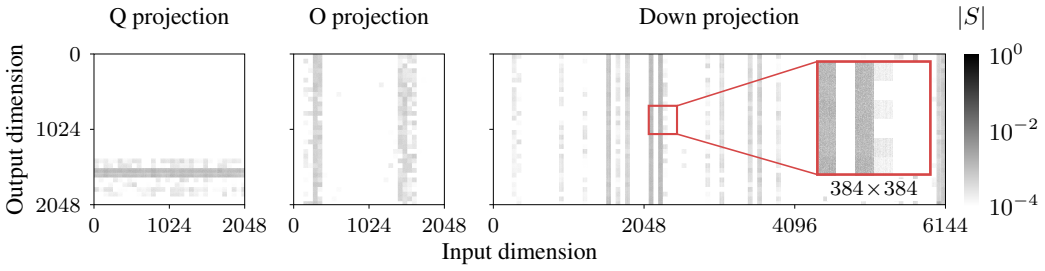


Figure 18: Learned sparse components S in layer 16 of Qwen3-1.7B trained with block-64×64 S-SALAAD. Columns show the Q, O, and down projections. The down-projection inset enlarges the red-boxed 384×384 region on the same color scale, showing 6×6 native blocks.

E.3 PAIRED SPARSITY PATTERNS

Beyond the structure within individual projections, the learned sparse components exhibit paired patterns between Q and K and between V and O within the same layer, as illustrated in Figure 19. The figure shows the sparse structure of a randomly selected layer of Llama-1B, with its five rows corresponding to unstructured sparsity (1×1) and four block sizes.

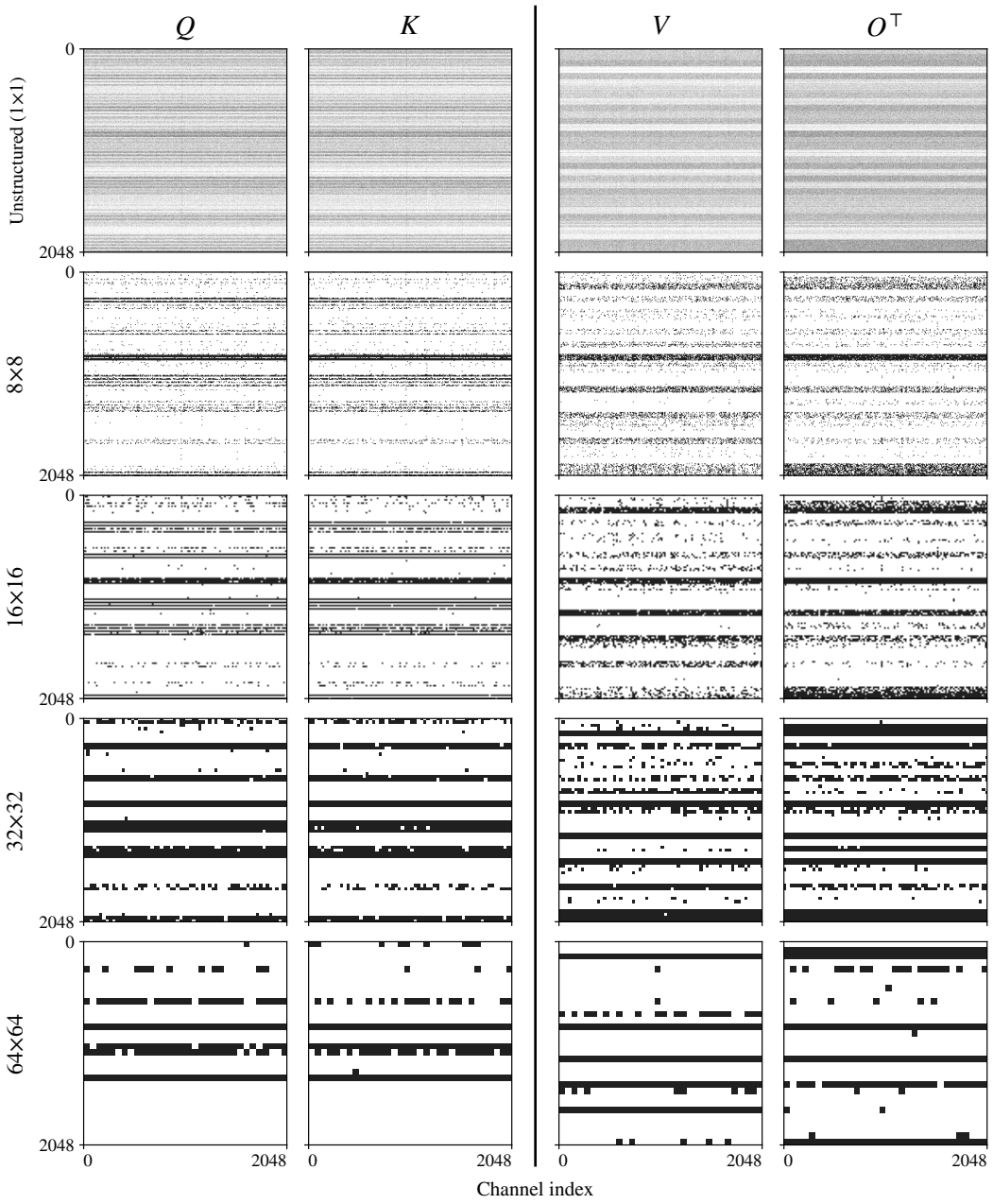


Figure 19: Paired sparsity patterns in a randomly selected layer of Llama-1B. Rows show unstructured sparsity (1×1) followed by four block sizes. Dark cells are nonzero and white cells are exactly zero. The O component is transposed so that its head slices align with those of V along rows.

Paired sparsity within a layer. In each row of Figure 19, nonzero entries in Q and K concentrate at similar row positions, while the nonzero regions in V and $O^\top$ also show correspondence across matching head slices. Each layer has 32 attention heads of dimension 64. Q/K pairing compares

the output-row slices belonging to the same head, while V/O pairing compares a V output-row slice with the corresponding O input-column slice. This consistency concerns the distribution of nonzero entries over heads, it does not require identical block supports within paired slices.

To examine whether this pattern extends to other layers, we analyze all 24 layers of the five checkpoints. For each attention projection $\mathrm{P} \in \{\mathrm{Q}, \mathrm{K}, \mathrm{V}, \mathrm{O}\}$, let $d_{\mathrm{P},h}$ denote the fraction of nonzero entries in the slice of its sparse component corresponding to head h. For each pair $(\mathrm{A}_1, \mathrm{A}_2) \in \{(\mathrm{Q}, \mathrm{K}), (\mathrm{V}, \mathrm{O})\}$, the Pearson correlation within a layer is

$$r_{\mathrm{A}_1,\mathrm{A}_2} = \frac{\sum_{h=1}^{H}(d_{\mathrm{A}_1,h} - \bar{d}_{\mathrm{A}_1})(d_{\mathrm{A}_2,h} - \bar{d}_{\mathrm{A}_2})}{\sqrt{\sum_{h=1}^{H}(d_{\mathrm{A}_1,h} - \bar{d}_{\mathrm{A}_1})^2}\sqrt{\sum_{h=1}^{H}(d_{\mathrm{A}_2,h} - \bar{d}_{\mathrm{A}_2})^2}},$$

where $H = 32$ and $\bar{d}_{\mathrm{P}} = H^{-1}\sum_{h=1}^{H} d_{\mathrm{P},h}$. Layers in which either projection has an entirely zero sparse component are excluded from the statistics for that pair, since the Pearson correlation is undefined. Table 9 reports the correlations across valid layers as mean $\pm$ standard deviation, showing strong average Q/K and V/O correlations at every block size. These results indicate that the sparse components of Q/K and V/O tend to concentrate on corresponding heads within the same layer.

Table 9: Mean $\pm$ sample standard deviation of layer-wise Pearson correlations of head nonzero fractions. Parentheses give the number n of valid layers for each pair, excluding layers in which either projection has an entirely zero sparse component.

Block size	Q/K correlation (layers)	V/O correlation (layers)
Unstructured (1×1)	0.971 ± 0.096 (24)	0.976 ± 0.024 (24)
8×8	0.986 ± 0.025 (23)	0.962 ± 0.051 (24)
16×16	0.979 ± 0.039 (23)	0.948 ± 0.076 (24)
32×32	0.980 ± 0.018 (21)	0.937 ± 0.078 (24)
64×64	0.967 ± 0.022 (21)	0.909 ± 0.127 (24)

Consistency of dominant heads across granularities. The rows of Figure 19 show that the main concentrations of nonzero entries occupy similar head positions across sparsity granularities. To quantify this consistency, we compare the same projection $\mathrm{P} \in \{\mathrm{Q}, \mathrm{K}, \mathrm{V}, \mathrm{O}\}$ across granularities. Let $d^{(b)}_{\mathrm{P},h}$ denote the nonzero fraction of head h at granularity $b\times b$, where $b \in \{1, 8, 16, 32, 64\}$. At each granularity, we select the four heads with the largest nonzero fractions and denote their index set by $T^{(b)}_{\mathrm{P}}$. For two granularities b and b', the number of shared heads is $|T^{(b)}_{\mathrm{P}} \cap T^{(b')}_{\mathrm{P}}|$. We include a layer comparison only when both granularities have at least four nonzero head slices and no tie between the fourth and fifth positions.

Table 10 reports the mean and sample standard deviation across valid layers separately for each of the ten granularity pairs and each projection.

Table 10: Number of shared head indices among the four heads with the largest nonzero fractions, comparing the same projection across each pair of sparsity granularities. Parentheses give the number of valid layers.

Granularity pair	Q	K	V	O
1×1 / 8×8	3.04 ± 0.77 (23)	2.73 ± 0.88 (22)	3.08 ± 0.72 (24)	2.88 ± 0.61 (24)
1×1 / 16×16	3.11 ± 0.88 (19)	2.72 ± 0.83 (18)	2.92 ± 0.78 (24)	2.92 ± 0.72 (24)
1×1 / 32×32	3.29 ± 0.77 (17)	3.00 ± 0.61 (17)	3.00 ± 0.63 (16)	2.91 ± 0.83 (11)
1×1 / 64×64	3.29 ± 0.73 (14)	2.75 ± 0.68 (16)	3.12 ± 0.81 (16)	2.93 ± 0.73 (14)
8×8 / 16×16	3.37 ± 0.60 (19)	3.06 ± 0.73 (18)	3.25 ± 0.61 (24)	3.25 ± 0.68 (24)
8×8 / 32×32	3.12 ± 0.60 (17)	2.94 ± 0.66 (17)	3.38 ± 0.50 (16)	3.27 ± 0.65 (11)
8×8 / 64×64	3.36 ± 0.63 (14)	3.06 ± 0.68 (16)	3.38 ± 0.62 (16)	3.29 ± 0.73 (14)
16×16 / 32×32	3.44 ± 0.81 (16)	3.33 ± 0.90 (15)	3.38 ± 0.50 (16)	3.36 ± 0.50 (11)
16×16 / 64×64	3.58 ± 0.67 (12)	3.15 ± 0.69 (13)	3.25 ± 0.58 (16)	3.50 ± 0.65 (14)
32×32 / 64×64	3.31 ± 0.75 (13)	3.20 ± 0.68 (15)	3.31 ± 0.63 (13)	3.22 ± 0.44 (9)

This consistency concerns the overlap of dominant heads and does not imply identical sets of nonzero heads or identical connections within them. Among the eligible layers, the dominant head locations within each projection are broadly consistent across the examined sparsity granularities.

E.4 Effective-Rank and Sparsity Ratios

Figures 20, 21, and 22 show how the effective-rank ratio and sparsity of sampled projections progressively approach their prescribed targets during training. We use a sparsity target of $\widehat{\Upsilon} = 0.95$ and an effective-rank ratio target of $\widehat{\Gamma} = 0.15$. For the $2:4$ pattern, the fixed 50% sparsity level gives $\widehat{\Upsilon} = 0.5$, and we correspondingly lower the rank target to $\widehat{\Gamma} = 0.03$.

Figure 20 compares block-16×16, row-/column-wise, and $2:4$ sparsity on Llama-350M. Figure 21 compares four block granularities on Llama-1B. Figure 22 shows the corresponding trajectories for three randomly selected projections during Qwen3-1.7B distillation with block-64×64 S-SALAAD.

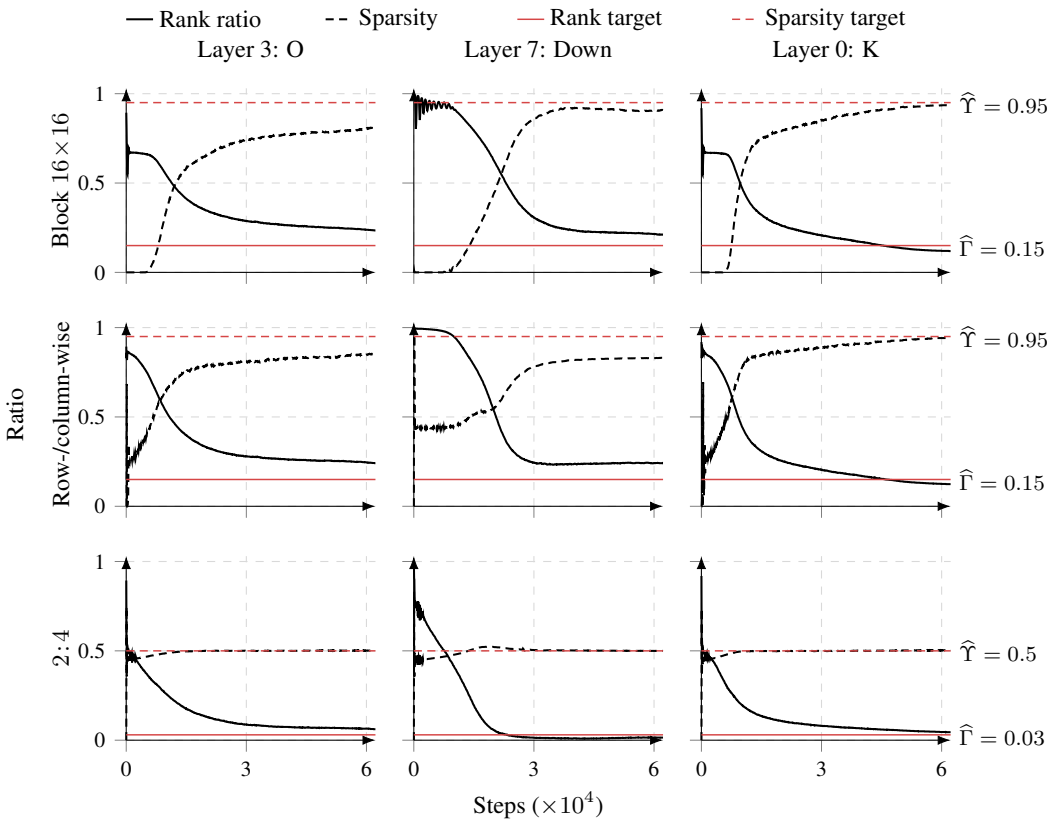


Figure 20: Effective-rank ratio and sparsity during Llama-350M pretraining. Rows correspond to block-16×16, row-/column-wise, and $2:4$ sparsity; columns show the O projection in layer 3, down projection in layer 7, and K projection in layer 0. Black solid and dashed curves show the rank ratio and sparsity (the zero fraction), respectively; red lines show the corresponding controller targets. The statistics are unsmoothed.

Across the evaluated sparsity patterns, model sizes, and architectures, the weight components progressively approach the prescribed structure, demonstrating the effectiveness of S-SALAAD's structure induction. The trajectories also reveal distinct convergence behavior across layers and projection types.

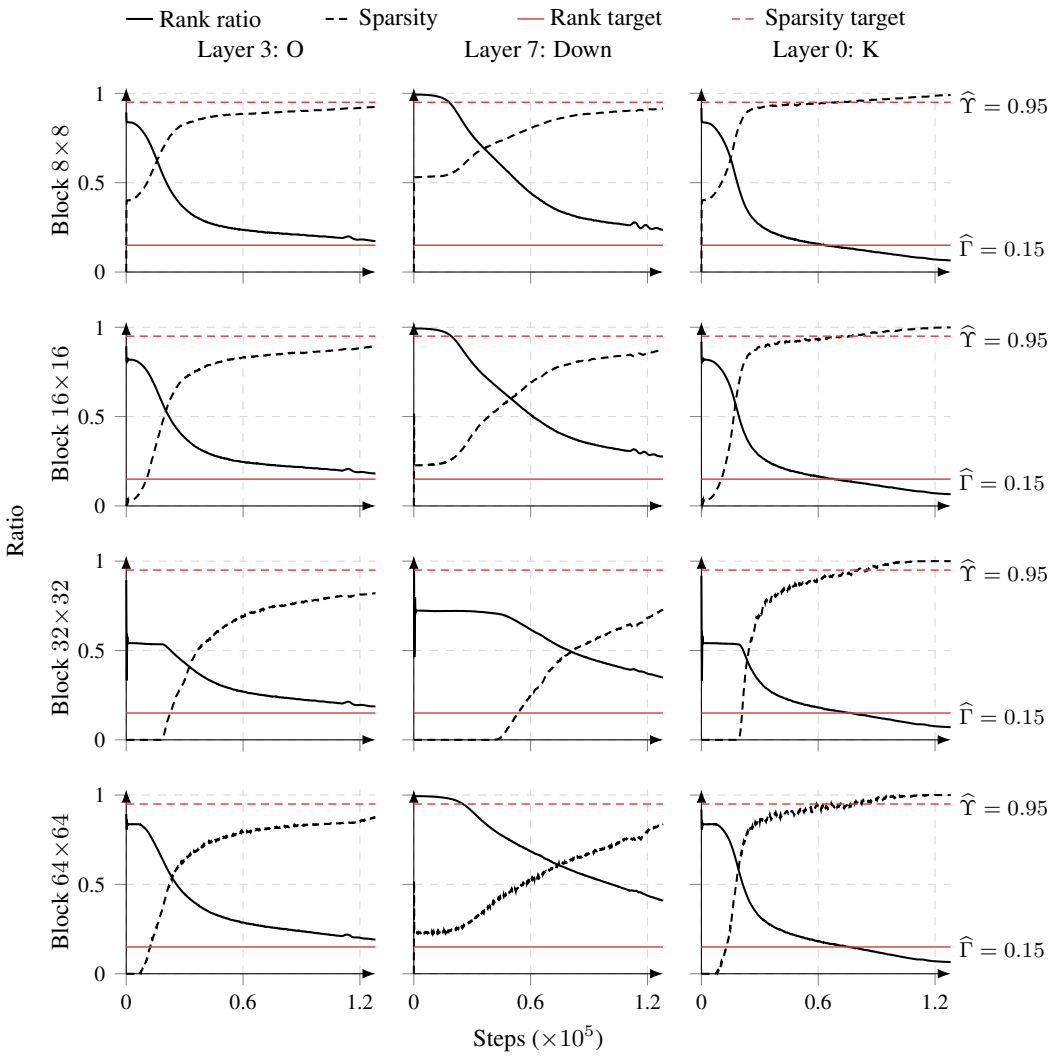


Figure 21: Effective-rank ratio and sparsity during Llama-1B pretraining. Rows correspond to block sizes 8, 16, 32, and 64; columns show the O projection in layer 3, down projection in layer 7, and K projection in layer 0. Black solid and dashed curves show the rank ratio and sparsity, respectively; red lines indicate their targets. The statistics are unsmoothed.

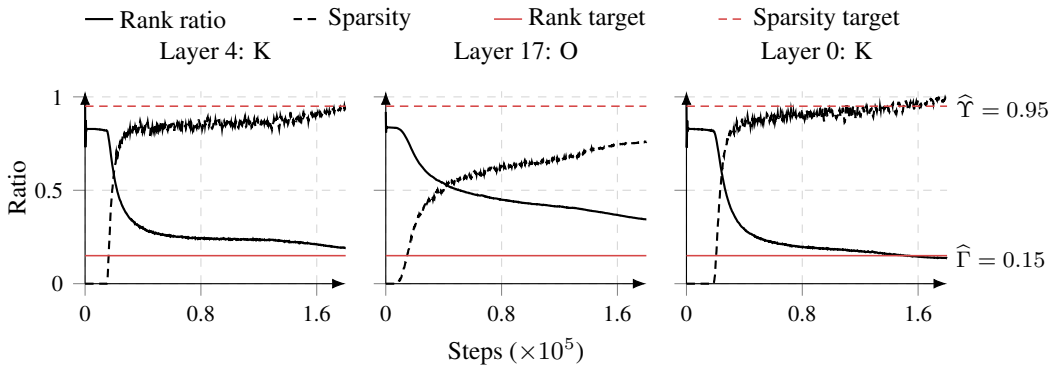
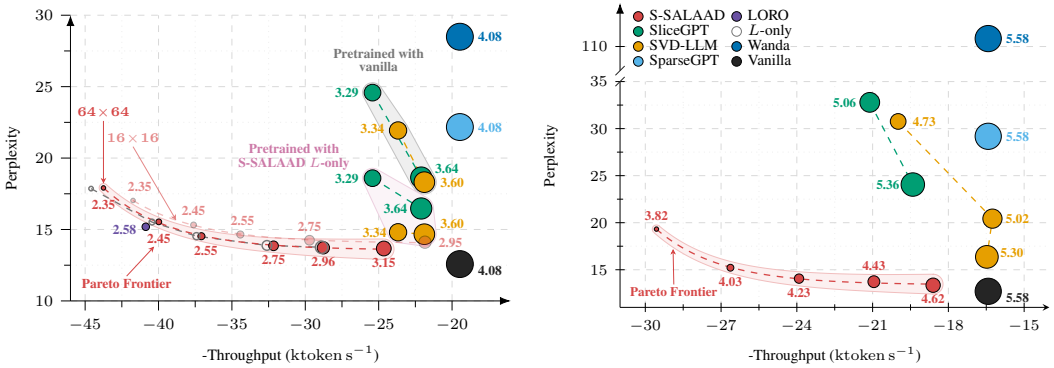


Figure 22: Effective-rank ratio and sparsity during Qwen3-1.7B distillation with block-64×64 S-SALAAD. Columns show the randomly selected K projection in layer 4, O projection in layer 17, and K projection in layer 0. Black solid and dashed curves show the rank ratio and sparsity, respectively; red lines indicate their targets. The statistics are unsmoothed.

F Additional Results

F.1 Pareto Frontiers on Other Hardware



(a) Llama-1B.

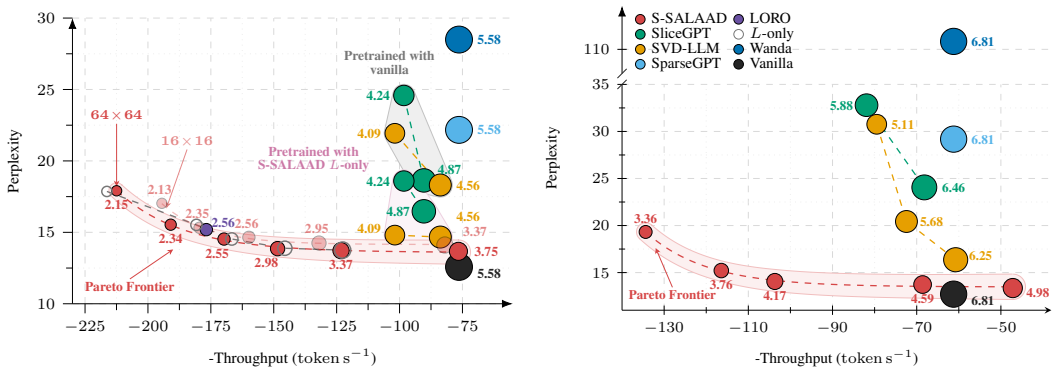
(b) Qwen3-1.7B.

Figure 23: Perplexity against throughput on an RTX 2080 Ti; marker radius linearly encodes peak GPU memory usage, with values annotated in GB. Lighter red denotes the 16×16 variant. Both panels share the legend in the right panel.

Figures 23 and 24 extend the deployment evaluation to an RTX 2080 Ti and 4× AMD EPYC 7662 CPUs. Relative to vanilla inference, block-64×64 S-SALAAD at a 500M budget reduces peak GPU memory usage by 42.4% and achieves 2.25× throughput on Llama-1B with the RTX 2080 Ti (Figure 23a). For Qwen3-1.7B at 800M, the corresponding gains are 31.5% lower peak GPU memory usage and 1.80× throughput (Figure 23b). On CPU, these compressed variants reduce peak RAM usage by 61.5% with 2.78× throughput for Llama-1B (Figure 24a), and by 50.6% with 2.20× throughput for Qwen3-1.7B (Figure 24b). These results demonstrate that S-SALAAD's memory and inference gains extend across the evaluated hardware platforms. Detailed numerical results are provided in Appendix H.

The deployment conclusions of Sections 5.1 and 5.2 also extend to these settings.

Synergy with post-training compression. Applying SliceGPT and SVD-LLM to L-only checkpoints pretrained with S-SALAAD yields lower perplexity at the same inference cost than applying them to vanilla-pretrained checkpoints. This benefit persists on both GPU and CPU (Figures 23a and 24a), demonstrating that the learned compressibility remains useful across hardware.



(a) Llama-1B. (b) Qwen3-1.7B.

Figure 24: Perplexity against throughput on 4× AMD EPYC 7662 CPUs (batch size 8, sequence length 64); marker radius is proportional to peak RAM usage, with values annotated in GB. Lighter red denotes the 16×16 variant. Both panels share the legend in the right panel.

Block granularity and deployment trade-offs. On Llama-1B, the 16×16 variant achieves lower perplexity than L-only under tight compression budgets (Figure 6b). Its additional inference overhead, however, limits the deployment gains on both GPU and CPU (Figures 23a and 24a). Increasing the block size reduces this overhead, allowing 64×64 to closely track the L-only perplexity–memory–latency trade-offs on the RTX 2080 Ti and improve on them at CPU operating points. These trends support the conclusion that retaining S can improve deployment trade-offs beyond L-only when block granularity is matched to the target hardware, as further illustrated in Appendix G.

F.2 Qwen Inference Under a Fixed Memory Cap

Under a constrained GPU memory budget, S-SALAAD's memory savings also enable larger inference batches, providing additional throughput gains. Figure 25 evaluates Qwen3-1.7B on the Quadro RTX 6000 under a fixed 3.4 GB peak-memory cap, with the batch size increased for each model within this limit. At a 1B parameter budget, S-SALAAD achieves $2.48\times$ the throughput of dense inference, with perplexity increasing from 12.71 to 14.06. These results show how elastic compression allows limited GPU memory to support higher throughput while offering a choice of model quality.

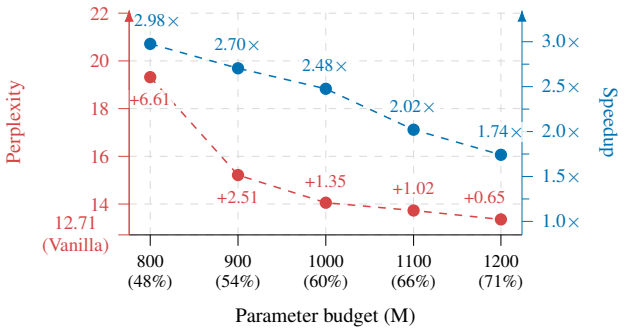
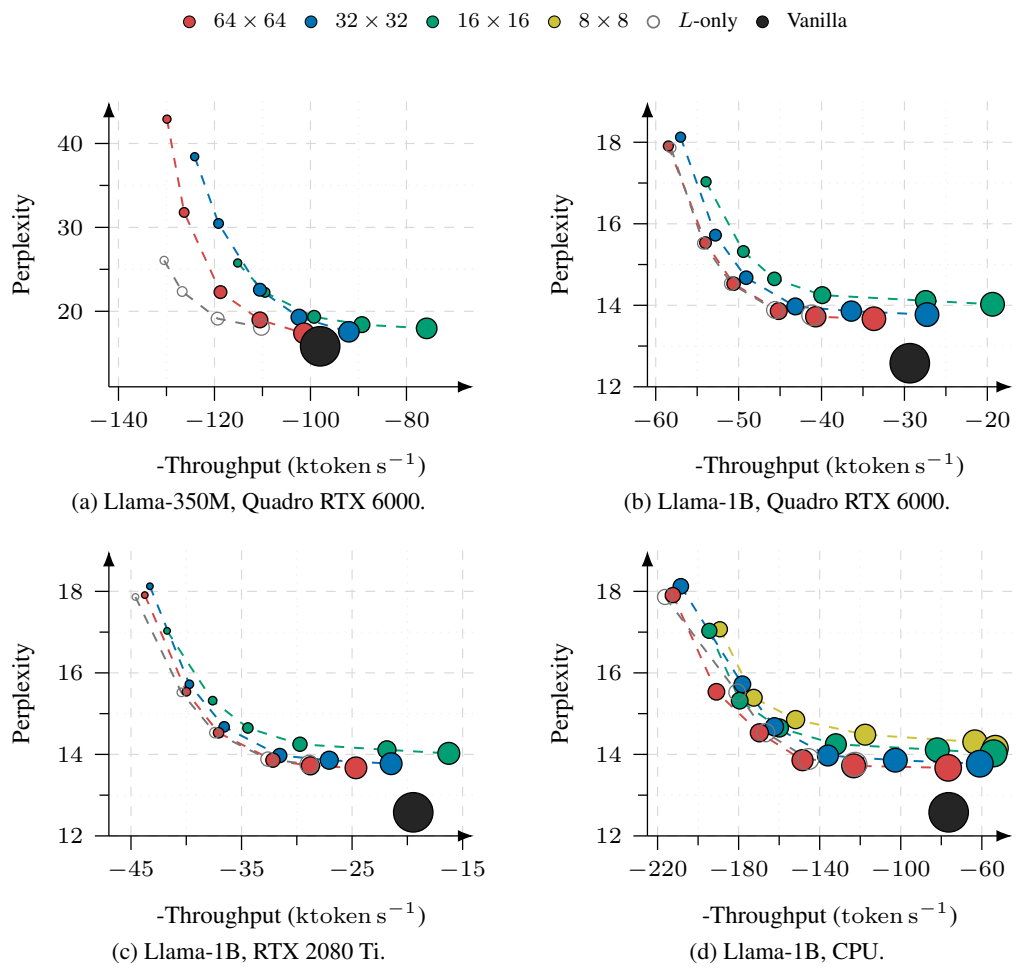


Figure 25: Compressed Qwen3-1.7B: speedup and perplexity under a 3.4 GB peak-memory cap. We increase the batch size until the cap.

G Block Granularity and the Pareto Frontier

Figure 26 extends the analysis in Section 5.2 by comparing the deployment trade-offs of different block granularities and the L-only special case across model scales and hardware.



(a) Llama-350M, Quadro RTX 6000.

(b) Llama-1B, Quadro RTX 6000.

(c) Llama-1B, RTX 2080 Ti.

(d) Llama-1B, CPU.

Figure 26: Perplexity against throughput on Llama for different block granularities, including the L-only baseline. Marker radius encodes peak GPU memory with min–max scaling or peak RAM with proportional scaling, per panel.

Finer blocks better preserve quality under tight compression budgets, but their additional inference overhead limits the deployment gains. On Llama-350M, coarser blocks incur a pronounced quality penalty despite their higher throughput (Figure 26a). On Llama-1B, the coarsest blocks retain competitive quality and closely track the L-only deployment trade-offs on both GPUs (Figures 26b and 26c). On CPU, the same granularity also provides operating points that improve on L-only (Figure 26d). Together with Figure 6, these trends indicate that larger models tolerate coarser structure, enabling more efficient inference while preserving elastic compressibility. They also support the main-text conclusion that retaining S can improve the perplexity–memory–latency trade-offs beyond L-only when block granularity is matched to the model and deployment hardware. Numerical results for Figure 26 are provided in Appendix H.

H Numerical Results

Tables 11, 12, and 13 list all operating points displayed in Figures 4, 23, and 24, respectively. Table 14 lists the operating points in Figure 26 not already reported in Tables 11, 12, and 13.

Table 11: Numerical results for Figure 4 on a Quadro RTX 6000. FP16; batch size × sequence length is 128×256 for Llama and 32×512 for Qwen. Arrows indicate preferred directions. Within each model, **bold** and underlined values mark the best and second-best distinct values per metric across the listed settings, at the displayed precision. Blue rows denote S-SALAAD, including its L-only special case. For the Llama baselines, (V) and (L) denote compression of a vanilla-pretrained and of an S-SALAAD L-only checkpoint, respectively.

Method	**Budget** (M)	**Params.**↓ (M)	**Perplexity**↓	**Peak mem.**↓ (GB)	**Latency**↓ (ms)	**Throughput**↑ (ktoken s^{-1})
			Llama-1B			
Vanilla	–	1364.8	**12.58**	4.08	1116	29.35
S-SALAAD (64×64)	500	**500.0**	17.91	**2.35**	**560**	**58.47**
	550	550.0	15.54	2.45	607	53.99
	600	600.0	14.53	2.55	647	50.61
	700	700.0	13.86	2.75	726	45.16
	800	800.0	13.72	2.96	805	40.70
	900	895.0	13.67	3.15	973	33.68
S-SALAAD (16×16)	500	**500.0**	17.03	**2.35**	608	53.93
	550	550.0	15.32	2.45	663	49.43
	600	600.0	14.65	2.55	717	45.69
	700	700.0	14.25	2.75	821	39.91
	800	800.0	14.11	2.95	1194	27.44
S-SALAAD (L-only)	500	**500.0**	17.86	**2.35**	563	58.17
	550	550.0	15.52	2.45	604	54.25
	600	600.0	14.53	2.55	644	50.90
	700	700.0	13.89	2.75	718	45.61
	800	799.9	13.76	2.96	796	41.15
LORO	–	616.1	15.19	2.58	602	54.45
SliceGPT (V)	–	1033.4	24.58	3.29	888	36.89
	–	1189.1	18.64	3.64	1021	32.11
SliceGPT (L)	–	1033.4	18.59	3.29	888	36.89
	–	1189.1	16.47	3.64	1021	32.11
SVD-LLM (V)	–	994.9	21.93	3.34	965	33.95
	–	1118.2	18.31	3.60	1055	31.07
SVD-LLM (L)	–	994.9	14.81	3.34	965	33.95
	–	1118.2	14.67	3.60	1055	31.07
Wanda	–	748.3	28.49	4.08	1116	29.35
SparseGPT	–	748.3	22.17	4.08	1116	29.35
			Qwen3-1.7B			
Vanilla	–	1677.8	**12.71**	5.58	677	24.19
S-SALAAD (64×64)	800	**800.0**	19.32	**3.82**	**418**	**39.22**
	900	900.0	15.21	4.03	455	36.03
	1000	1000.0	14.06	4.23	493	33.23
	1100	1100.0	13.73	4.43	578	28.35
	1200	1199.5	13.36	4.62	671	24.42
SliceGPT	–	1449.1	32.79	5.06	544	30.10
	–	1591.8	24.05	5.36	590	27.75
SVD-LLM	–	1254.6	30.76	4.73	585	28.02
	–	1395.4	20.43	5.02	653	25.09
	–	1536.4	16.37	5.30	701	23.38
Wanda	–	973.2	113.29	5.58	677	24.19
SparseGPT	–	973.2	29.18	5.58	677	24.19

Table 12: Numerical results for Figure 23 on an RTX 2080 Ti. FP16; batch size $\times$ sequence length is 128×256 for Llama and 32×512 for Qwen. Arrows indicate preferred directions. Within each model, **bold** and underlined values mark the best and second-best distinct values per metric across the listed settings, at the displayed precision. Blue rows denote S-SALAAD, including its L-only special case. For the Llama baselines, (V) and (L) denote compression of a vanilla-pretrained and of an S-SALAAD L-only checkpoint, respectively.

Method	**Budget** (M)	**Params.**↓ (M)	**Perplexity**↓	**Peak mem.**↓ (GB)	**Latency**↓ (ms)	**Throughput**↑ (ktoken s^{-1})
			Llama-1B			
Vanilla	–	1364.8	**12.58**	4.08	1683	19.47
S-SALAAD (64×64)	500	**500.0**	17.91	**2.35**	749	43.75
	550	550.0	15.54	2.45	820	39.98
	600	600.0	14.53	2.55	884	37.08
	700	700.0	13.86	2.75	1019	32.16
	800	800.0	13.72	2.96	1139	28.77
	900	895.0	13.67	3.15	1329	24.65
S-SALAAD (16×16)	500	**500.0**	17.03	**2.35**	785	41.74
	550	550.0	15.32	2.45	871	37.61
	600	600.0	14.65	2.55	952	34.42
	700	700.0	14.25	2.75	1103	29.71
	800	800.0	14.11	2.95	1500	21.85
S-SALAAD (L-only)	500	**500.0**	17.86	**2.35**	**735**	**44.60**
	550	550.0	15.52	2.45	810	40.45
	600	600.0	14.53	2.55	876	37.42
	700	700.0	13.89	2.75	1005	32.61
	800	799.9	13.76	2.96	1134	28.89
LORO	–	616.1	15.19	2.58	802	40.87
SliceGPT (V)	–	1033.4	24.58	3.29	1289	25.42
	–	1189.1	18.64	3.64	1483	22.10
SliceGPT (L)	–	1033.4	18.59	3.29	1289	25.42
	–	1189.1	16.47	3.64	1483	22.10
SVD-LLM (V)	–	994.9	21.93	3.34	1384	23.68
	–	1118.2	18.31	3.60	1497	21.89
SVD-LLM (L)	–	994.9	14.81	3.34	1384	23.68
	–	1118.2	14.67	3.60	1497	21.89
Wanda	–	748.3	28.49	4.08	1683	19.47
SparseGPT	–	748.3	22.17	4.08	1683	19.47
			Qwen3-1.7B			
Vanilla	–	1677.8	**12.71**	5.58	998	16.42
S-SALAAD (64×64)	800	**800.0**	19.32	**3.82**	**554**	**29.56**
	900	900.0	15.21	4.03	615	26.63
	1000	1000.0	14.06	4.23	685	23.91
	1100	1100.0	13.73	4.43	782	20.95
	1200	1199.5	13.36	4.62	881	18.60
SliceGPT	–	1449.1	32.79	5.06	776	21.11
	–	1591.8	24.05	5.36	844	19.41
SVD-LLM	–	1254.6	30.76	4.73	820	19.99
	–	1395.4	20.43	5.02	1008	16.26
	–	1536.4	16.37	5.30	994	16.48
Wanda	–	973.2	113.29	5.58	998	16.42
SparseGPT	–	973.2	29.18	5.58	998	16.42

Table 13: Numerical results for Figure 24 on 4× AMD EPYC 7662 CPUs. FP32; batch size 8 and sequence length 64 for both models. Arrows indicate preferred directions. Within each model, **bold** and underlined values mark the best and second-best distinct values per metric across the listed settings, at the displayed precision. Blue rows denote S-SALAAD, including its L-only special case. For the Llama baselines, (V) and (L) denote compression of a vanilla-pretrained and of an S-SALAAD L-only checkpoint, respectively.

Method	**Budget** (M)	**Params.**↓ (M)	**Perplexity**↓	**Peak mem.**↓ (GB)	**Latency**↓ (ms)	**Throughput**↑ (token s^{-1})
			Llama-1B			
Vanilla	–	1364.8	**12.58**	5.58	6705	76.36
S-SALAAD (64×64)	500	**500.0**	17.91	<u>2.15</u>	<u>2410</u>	<u>212.46</u>
	550	<u>550.0</u>	15.54	2.34	2681	190.96
	600	600.0	14.53	2.55	3016	169.77
	700	700.0	13.86	2.98	3449	148.47
	800	800.0	13.72	3.37	4154	123.26
	900	895.0	<u>13.67</u>	3.75	6685	76.59
S-SALAAD (16×16)	500	**500.0**	17.03	**2.13**	2633	194.45
	550	<u>550.0</u>	15.32	2.35	2853	179.48
	600	600.0	14.65	2.56	3202	159.88
	700	700.0	14.25	2.95	3878	132.02
	800	800.0	14.11	3.37	6245	81.98
S-SALAAD (L-only)	500	**500.0**	17.86	**2.13**	**2367**	**216.32**
	550	<u>550.0</u>	15.52	2.34	2833	180.72
	600	600.0	14.53	2.56	3072	166.67
	700	700.0	13.89	2.95	3517	145.58
	800	799.9	13.76	3.37	4176	122.60
LORO	–	616.1	15.19	2.56	2896	176.79
SliceGPT (V)	–	1033.4	24.58	4.24	5209	98.29
	–	1189.1	18.64	4.87	5667	90.34
SliceGPT (L)	–	1033.4	18.59	4.24	5209	98.29
	–	1189.1	16.47	4.87	5667	90.34
SVD-LLM (V)	–	994.9	21.93	4.09	5026	101.86
	–	1118.2	18.31	4.56	6110	83.79
SVD-LLM (L)	–	994.9	14.81	4.09	5026	101.86
	–	1118.2	14.67	4.56	6110	83.79
Wanda	–	748.3	28.49	5.58	6705	76.36
SparseGPT	–	748.3	22.17	5.58	6705	76.36
			Qwen3-1.7B			
Vanilla	–	1677.8	**12.71**	6.81	8363	61.22
S-SALAAD (64×64)	800	**800.0**	19.32	**3.36**	**3810**	**134.40**
	900	<u>900.0</u>	15.21	<u>3.76</u>	<u>4398</u>	<u>116.42</u>
	1000	1000.0	14.06	4.17	4937	103.70
	1100	1100.0	13.73	4.59	7464	68.60
	1200	1199.5	<u>13.36</u>	4.98	10854	47.17
SliceGPT	–	1449.1	32.79	5.88	6250	81.92
	–	1591.8	24.05	6.46	7509	68.19
SVD-LLM	–	1254.6	30.76	5.11	6441	79.50
	–	1395.4	20.43	5.68	7068	72.43
	–	1536.4	16.37	6.25	8429	60.74
Wanda	–	973.2	113.29	6.81	8363	61.22
SparseGPT	–	973.2	29.18	6.81	8363	61.22

Table 14: Additional numerical results for Figure 26. Throughput units appear in the group headings. Arrows indicate preferred directions. **Bold** and underlined values mark the best and second-best distinct values at the displayed precision within each complete plotted panel, including entries referenced in Tables 11–13.

Method	Budget (M)	Params.↓ (M)	Perplexity↓	Peak mem.↓ (GB)	Latency↓ (ms)	Throughput↑
Llama-350M, Quadro RTX 6000 (ktoken s^{-1})						
Vanilla	–	374.1	**15.83**	0.92	84	97.98
S-SALAAD (64×64)	170	**170.0**	42.90	**0.52**	**63**	129.90
	180	180.0	31.78	0.54	65	126.32
	200	200.0	22.29	0.58	69	118.75
	220	220.0	18.98	0.62	74	110.53
	250	250.0	17.38	0.68	81	101.35
S-SALAAD (32×32)	170	**170.0**	38.42	**0.52**	66	124.13
	180	180.0	30.46	0.54	69	119.15
	200	200.0	22.57	0.58	74	110.53
	220	220.0	19.30	0.62	80	102.39
	250	250.0	17.56	0.68	89	91.99
S-SALAAD (16×16)	170	**170.0**	25.76	**0.52**	71	115.16
	180	180.0	22.25	0.54	75	109.46
	200	200.0	19.34	0.58	83	99.23
	220	220.0	18.41	0.62	92	89.28
	250	250.0	17.96	0.68	108	75.78
S-SALAAD (L-only)	170	**170.0**	26.06	**0.52**	**63**	**130.49**
	180	180.0	22.36	0.54	65	126.73
	200	200.0	19.13	0.58	69	119.34
	220	220.0	18.09	0.62	74	110.23
Llama-1B, Quadro RTX 6000 (ktoken s^{-1})						
S-SALAAD (32×32)	500	**500.0**	18.13	**2.35**	575	57.00
	550	550.0	15.72	2.45	621	52.80
	600	600.0	14.68	2.55	668	49.09
	700	700.0	13.97	2.75	759	43.16
	800	800.0	13.86	2.95	900	36.41
	900	899.5	13.77	3.15	1201	27.28
S-SALAAD (16×16)	900	900.0	14.03	3.15	1691	19.37
Llama-1B, RTX 2080 Ti (ktoken s^{-1})						
S-SALAAD (32×32)	500	**500.0**	18.13	**2.35**	757	43.29
	550	550.0	15.72	2.45	825	39.72
	600	600.0	14.68	2.55	896	36.57
	700	700.0	13.97	2.75	1039	31.54
	800	800.0	13.86	2.95	1211	27.07
	900	899.5	13.77	3.15	1526	21.47
S-SALAAD (16×16)	900	900.0	14.03	3.15	2016	16.25
Llama-1B, AMD EPYC 7662 (token s^{-1})						
S-SALAAD (32×32)	500	**500.0**	18.13	2.17	2455	208.51
	550	550.0	15.72	2.35	2874	178.16
	600	600.0	14.68	2.57	3154	162.34
	700	700.0	13.97	2.95	3766	135.94
	800	800.0	13.86	3.35	4987	102.67
	900	899.5	13.77	3.77	8394	60.99
S-SALAAD (16×16)	900	900.0	14.03	3.81	9443	54.22
S-SALAAD (8×8)	500	**500.0**	17.07	2.14	2705	189.25
	550	550.0	15.39	2.37	2968	172.53
	600	600.0	14.85	2.56	3371	151.90
	700	700.0	14.48	2.95	4356	117.55
	800	800.0	14.31	3.39	8070	63.44
	900	866.0	14.15	3.64	9596	53.35