Numerical Pruning for Efficient Autoregressive Models

Xuan Shen[1*], Zhao Song[2], Yufa Zhou[3], Bo Chen[4], Jing Liu[5], Ruiyi Zhang[2], Ryan A. Rossi[2], Hao Tan[2], Tong Yu[2], Xiang Chen[2], Yufan Zhou[2], Tong Sun[2], Pu Zhao[1], Yanzhi Wang[1†], Jiuxiang Gu[2†]

[1]Northeastern University
[2]Adobe Research
[3]University of Pennsylvania
[4]Middle Tennessee State University
[5]Monash University
{shen.xu, p.zhao, yanz.wang}@northeastern.edu, jigu@adobe.com

Abstract

Transformers have emerged as the leading architecture in deep learning, proving to be versatile and highly effective across diverse domains beyond language and image processing. However, their impressive performance often incurs high computational costs due to their substantial model size. This paper focuses on compressing decoder-only transformer-based autoregressive models through structural weight pruning to improve the model efficiency while preserving performance for both language and image generation tasks. Specifically, we propose a training-free pruning method that calculates a numerical score with Newton's method for the Attention and MLP modules, respectively. Besides, we further propose another compensation algorithm to recover the pruned model for better performance. To verify the effectiveness of our method, we provide both theoretical support and extensive experiments. Our experiments show that our method achieves state-of-the-art performance with reduced memory usage and faster generation speeds on GPUs.

1 Introduction

Transformers have been dominant in generative models. This includes Large Language Models (LLMs) (Vaswani et al. 2017; Touvron et al. 2023b) for language generation, as well as recent autoregressive image generation models (Van Den Oord, Vinyals et al. 2017; Esser, Rombach, and Ommer 2021; Ramesh et al. 2021; Yu et al. 2022). Notably, models such as LlamaGen (Sun et al. 2024), which use image tokenizers to convert continuous images into discrete tokens, have demonstrated the ability to surpass diffusion models (Ho, Jain, and Abbeel 2020; Rombach et al. 2022) in image generation tasks. The "*next-token prediction*" paradigm demonstrates significant capabilities in addressing both language and image generation tasks, enabling solutions that mimic human-like conversational interactions (Achiam, Adler et al. 2023; Li et al. 2024a).

Recognizing the capabilities of large autoregressive models pioneering works (Frantar and Alistarh 2023; Sun et al. 2023; Ma, Fang, and Wang 2023; Ashkboos et al. 2024; Zhan et al. 2021; Zhao et al. 2024; Zhan et al. 2024b) have sought to compress these models to enhance their execution efficiency. Compared to irregular pruning methods, structural pruning offers a more efficient reduction in both computational and memory overhead (Jian et al. 2021; Gong et al. 2022, 2023). By maintaining a consistent and regular structure, it simplifies implementation, accelerates processing, and leads to more predictable resource savings (Kong et al. 2022, 2023). However, most of these efforts focus solely on language models and language-related research areas. Consequently, their methods are not readily applicable to image generation tasks because of the fundamental differences in data structure and computational requirements between language and image processing (Reed et al. 2016; Parmar et al. 2018; Lee et al. 2022; Shen et al. 2024a,c,b, 2023b,a; Li et al. 2023, 2024c,b). Therefore, it is crucial to explore the transformer architecture itself, rather than focusing on specific application models. This motivates us to develop a general method for compressing autoregressive models applicable to multiple kinds of generative tasks.

Additionally, the recovery of pruned models are crucial. Full-parameter retraining of large autoregressive models after pruning is often computationally prohibitive, making calibrations with a few samples a preferred approach. Previous work (Frantar and Alistarh 2023) employs the Optimal Brain Surgeon (OBS) technique (Hassibi, Stork, and Wolff 1993; LeCun, Denker, and Solla 1989) for weight updates during pruning. However, its heavy reliance on the approximation information increases sensitivity to noise and reduces robustness across different datasets. SliceGPT (Ashkboos et al. 2024) relies on a large number of samples for pruning and calibration, leading to overfitting on calibration data and limiting the generalization to other different datasets.

In this work, we present a novel structural pruning approach that leverages our proposed numerical score, combined with compensation techniques for performance recovery. We first calculate the numerical score for each layer through solving the optimal pruning mask for the minimization of pruning errors using the Newton's method. By ranking these numerical scores of all layers, we generate the globally pruning mask with the specified pruning ratio. Additionally, we introduce a compensation algorithm to recover

[*]Work done during internship at Adobe Research
[†]Corresponding Author

pruned models by updating the remaining weights to account for the loss caused by the pruned weights. We empirically evaluate our method using the LLaMA model family including LLaMA, LLaMA-2, and LLaMA-3 as representative LLMs and LlamaGen for image generation tasks. Experimental results show that our method outperforms other state-of-the-art approaches in both language and image generation tasks, validating the effectiveness of our proposed numerical score and compensation algorithm. Moreover, our method reduces GPU memory usage and accelerates generation without requiring any additional GPU-specific modifications. Our main contributions are summarized as follows,

- We propose a numerical score, derived from the numerical solution of the optimal mask for minimizing pruning errors with Newton's method.
- We propose a compensation algorithm for the reconstruction of the pruned model, further enhancing the task performance of the pruned model.
- Experimental results show that our method not only achieves state-of-the-art performance but also reduces memory usage and accelerates generation on GPUs.

2 Related Work

2.1 Compression for LLMs

The large number of parameters in LLMs motivates the need for pruning (Gong et al. 2020; Wu et al. 2022; Zhan et al. 2024a; Li et al. 2022; Zhang et al. 2022; Zhan et al. 2024c; Shen et al. 2024d) to improve efficiency. The work (Frantar and Alistarh 2023) introduces the Optimal Brain Surgeon (OBS) method (Hassibi, Stork, and Wolff 1993; LeCun, Denker, and Solla 1989) to compress the LLMs, which removes weights with minimal impact on the loss function. It then updates the remaining weights by utilizing the inverse of the Hessian matrix to mitigate errors caused by the pruning process. Unfortunately, this kind of pruning method is still irregular, meaning it does not lead to significant reductions in memory and computational requirements. Subsequent works, such as LLM-Pruner (Ma, Fang, and Wang 2023), SliceGPT (Ashkboos et al. 2024), and FLAP (An et al. 2023), propose structural pruning methods that effectively reduce memory usage and accelerate inference on GPUs. These methods offer significant advantages over irregular pruning by directly enhancing the utility and efficiency of the models. While autoregressive models excel in sequential data processing, such as text, the distinct nature of image data, where spatial relationships and pixel-level details are critical, demands different approaches. As a result, adapting these models to image generation introduces complexities that limit their scalability and effectiveness.

2.2 Autoregressive Models in Image Generation

Autoregressive models, initially renowned for their success with LLMs, have recently gained popularity in the image generation research area. Pioneering works (Van Den Oord, Vinyals et al. 2017; Esser, Rombach, and Ommer 2021) introduced image tokenizers that convert continuous images into discrete tokens. These tokenizers, which have been demonstrated to be effective by the following works (Ramesh et al. 2021; Yu et al. 2021, 2022), enable autoregressive models to generate image tokens using the next-token prediction approach. Recent work (Sun et al. 2024) delivers a series of image generation models with a new constructed image tokenizer. This research demonstrates the effectiveness of LLM frameworks in image generation tasks, validating their potential beyond traditional language applications. Additionally, the work (Li et al. 2024a) delves deeper into the continuous-valued domains of autoregressive models and removes the image tokenizers for image generation tasks. This work achieves stronger results while leveraging the speed advantage of sequence modeling, which further enhances the utilization and demonstrates the potential of autoregressive models in image generation tasks.

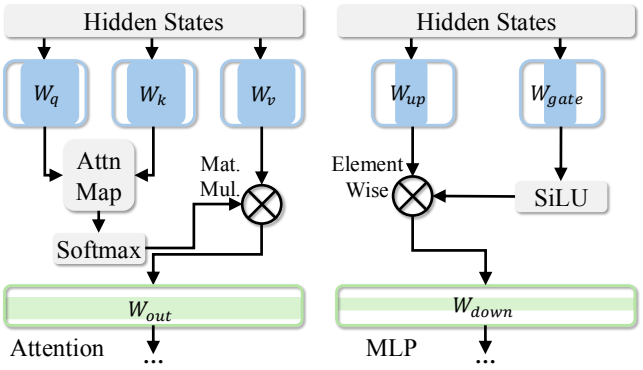

Figure 1: Pruning overview. Blue modules denote column pruning and green modules denote row pruning.

3 Methodology

3.1 Preliminary

Notations. We use $\mathbb{E}[\cdot]$ to denote the expectation. For two vectors $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$, we use $\langle x, y \rangle$ to denote the inner product between x, y, *i.e.*, $\langle x, y \rangle = \sum_{i=1}^n x_i y_i$. We use $\mathbf{1}_n$ to denote a length-n vector where all the entries are ones. We use $x_{i,j}$ to denote the j-th coordinate of $x_i \in \mathbb{R}^n$. We use $\|x\|_p$ to denote the ℓ_p norm of a vector $x \in \mathbb{R}^n$. For each $a, b \in \mathbb{R}^n$, we use $a \circ b \in \mathbb{R}^n$ to denote the vector where i-th entry is $(a \circ b)_i = a_i b_i$ for all $i \in [n]$. We use $\|A\|$ to denote the spectral norm for matrix A. For a square matrix A, we use $\mathrm{tr}[A]$ to denote the trace of A, *i.e.*, $\mathrm{tr}[A] = \sum_{i=1}^n A_{i,i}$.

Internal Computation Alignment. To maintain model interpretability and mitigate the risk of model drift, we ensure consistency in the internal computations of the Attention module. This approach is aligned with established methodologies in the literature (Vaswani et al. 2017; Yang et al. 2020). Considering the definition in this paper mainly focuses on $X \cdot W$ where $X \in \mathbb{R}^{N \times D}$ and $W \in \mathbb{R}^{D \times D'}$, we visualize the pruning strategy in Figure 1. In detail, we utilize the identical pruning mask (*i.e.*, pruning strategy) for the columns of weights associated with the query, key, and value, as well as for the rows of weights in the output projection. Meanwhile, we apply the same strategy to the MLP module, using column pruning for the up and gate weights,

and row pruning for the down projection weights. In this paper, we construct structural pruning metrics focusing on the output projection layers of Attention module and the down projection layers of MLP module.

3.2 Numerical Score

We define the weight as $W \in \mathbb{R}^{D \times D'}$, input as $X \in \mathbb{R}^{N \times D}$, and we denote the mask as $M \in \{0,1\}^D$. Additionally, we define the pruning ratio $\rho \in [0,1]$ as the ratio of the number of zeros to the total number of entries in pruning mask M.

Note that, when we apply the mask column by column, the mask M is a D-dimensional vector. Specifically, if $M_j = 0$ for $j \in [D]$, we prune the entire row for W, *i.e.*, $W_j = 0$, and if $M_j = 1$ we keep the original W_j.

To compute the numerical score, we explore the bound of the error (*i.e.*, difference) between the original weights and pruned weights. For the bound of the error, we first formulate the error for $i \in [D']$ as

$$\|XW_{*,i} - X(M \circ W_{*,i})\|_2. \tag{1}$$

Simplify further, for $i \in [D']$, Eq. (1) can be transformed into the following,

$$\|X((\mathbf{1}_D - M) \circ W_{*,i})\|_2.$$

In the above equation, the $\|\mathbf{1}_D - M\|_2$ denotes the number of zero entries in M, which is corresponding to the simply $\rho \cdot D$. Furthermore, assuming $\|X\| \leq R$, we demonstrate that the following Lemma 1 holds,

Lemma 1 (informal version of Lemma 9 at Appendix D.2)**.** *We show that for $i \in [D']$ we have*

$$\|XW_{*,i} - X(M \circ W_{*,i})\|_2 \leq \rho R \|W_{*,i}\|_2.$$

It is intuitive for us to minimize the error after establishing the error bound in Lemma 1. Thus, we examine each term. In the error bound of Lemma 1, the pruning ratio ρ is manually specified. We adopt the normalization for the input X, then the norm of normalized X is upper bounded by 1. Meanwhile, for $\|W_{*,i}\|_2$ term, it is the ℓ_2 norm for i-th column of weight W.

In order to minimize the error, we regulate both ρ and $\|W_{*,i}\|$. Then, we generalize the mask M from binary value to real value for the calculation of the numerical score. Meanwhile, we set one threshold which converts the real-valued mask back into a binary mask. For mask $M \in [0,1]^D$ and pruning ratio $\rho \in [0,1]$, the calculation of the numerical score is formulated as follows,

$$\begin{aligned} \arg\min_{M} \quad & \sum_{i \in [D']} \|XW_{*,i} - X(M \circ W_{*,i})\|_2, \\ \text{s.t.} \quad & \langle \mathbf{1}_D, M \rangle = (1-\rho)D. \end{aligned} \tag{2}$$

To better solve Eq. (2), we define the numerical score $z \in [0,1]^D$ and $r := (1-\rho)D \in [0, D]$. The equality constraint in Eq. (2) is then equivalent to $\langle \mathbf{1}_D, z \rangle - r = 0$.

Then, Eq. (2) becomes the minimization problem with the equality constraint. To efficiently solve such problem, we adopt the Newton' method (Bubeck et al. 2015). By turning the equality constraint into a penalty term for regularization, we further generate the following equivalent problem,

$$\begin{aligned} \arg\min_{z \in [0,1]^D} \quad & \frac{1}{2} \sum_{i \in [D']} \|XW_{*,i} - X(z \circ W_{*,i})\|_2^2 \\ & + \frac{1}{2} \lambda \cdot (\langle \mathbf{1}_D, z \rangle - r)^2, \end{aligned} \tag{3}$$

where $\lambda \in \mathbb{R}_+$ is the regularization parameter.

To explain how we solve this, we define the loss function for $i \in [D']$ as follows,

$$L(z)_i = \frac{1}{2} \|XW_{*,i} - X(z \circ W_{*,i})\|_2^2. \tag{4}$$

Meanwhile, for regularization term, we define as follows,

$$L_{\mathrm{reg}}(z) = \frac{1}{2} \lambda \cdot (\langle \mathbf{1}_D, z \rangle - r)^2. \tag{5}$$

Combining Lemma 12 and Lemma 13 at Appendix D.3, we compute the gradient of Eq. (4) and Eq. (5) as follows,

$$\begin{aligned} g = & \underbrace{((WW^\top) \circ (X^\top X))(z - \mathbf{1}_D)}_{\text{Gradient of } L(z)} \\ & + \underbrace{\lambda(\langle \mathbf{1}_D, z \rangle - r) \cdot \mathbf{1}_D}_{\text{Gradient of } L_{\mathrm{reg}}(z)}. \end{aligned} \tag{6}$$

Combining Lemma 14 and Lemma 15 at Appendix D.3, we compute the Hessian of Eq. (4) and Eq. (5) as follows,

$$H = \underbrace{(WW^\top) \circ (X^\top X)}_{\text{Hessian of } L(z)} + \underbrace{\lambda \cdot \mathbf{1}_{D \times D}}_{\text{Hessian of } L_{\mathrm{reg}}(z)}. \tag{7}$$

Subsequently, using Algorithm 1, we efficiently compute the optimal numerical z in $O(TD^3)$, where T represents the number of iterations for Newton's Method, typically around 50 in practice. Besides, we derive the following Theorem 2.

Theorem 2 (Mask optimization, informal version of Theorem 10 at Appendix D.3)**.** *If the following conditions hold:*

- *Let $W \in \mathbb{R}^{D \times D'}$, $X \in \mathbb{R}^{N \times D}$.*

Algorithm 1: Numerical Score with Newton's Method

procedure NUMERICALSCORE($X \in \mathbb{R}^{N \times D}, W \in \mathbb{R}^{D \times D'}, r \in [0, D], \lambda \in \mathbb{R}_+, T \in \mathbb{N}_+$) ▷ Theorem 2
  We choose the initial point z_0 such that $z_0 \in [0,1]^D$
  for $t = 0 \to T$ **do**
    $g_l \leftarrow ((WW^\top) \circ (X^\top X))(z - \mathbf{1}_D)$
    $g_r \leftarrow \lambda(\langle \mathbf{1}_D, z \rangle - r) \cdot \mathbf{1}_D$
    $g \leftarrow g_l + g_r \in \mathbb{R}^D$
    $H_l \leftarrow (WW^\top) \circ (X^\top X)$
    $H_r \leftarrow \lambda \cdot \mathbf{1}_{D \times D}$
    $H \leftarrow H_l + H_r \in \mathbb{R}^{D \times D}$
    $z_{t+1} \leftarrow z_t - H^{-1} g$
  end for
  $z \leftarrow z_{T+1}$
  return z
end procedure

- *Let* $z \in [0,1]^D$.
- *Let* $r \in [0, D]$ *denote the number of ones (it can be a fractional number).*
- *Let* $\lambda > 0$ *denote a regularization co-efficients.*
- *Assume* $\|X\| \leq R$.

There exists an algorithm (Algorithm 1) that can get the optimal z *in* $O(TD^3)$ *for Eq.* (3)*.*

3.3 Global Pruning

To ensure consistent head-level computation in the Attention module, given numerical scores $z^{\text{attn}} \in \mathbb{R}^D$, we group the scores of channels for h-th head to determine the importance score z_h^{head} of individual heads as follows,

$$z_h^{\text{head}} = \frac{1}{D_h} \cdot \sum_{i=h\cdot D_h}^{(h+1)\cdot D_h} z_i^{\text{attn}},$$

where D_h denotes the dimension of each head, $h \in [H]$ is the head index.

Unlike the Attention module, where heads work in parallel to capture various aspects of the input and their outputs are interdependent, the MLP module has a simpler structure with minimal interdependencies between its components. Thus, we retain the channel scores $z^{\text{mlp}} \in \mathbb{R}^D$ to guide the pruning process for MLP module.

To ensure a balanced pruning process that reflects the relative importance of each layer, we simultaneously sort the numerical scores across all layers to derive the globally pruning mask. Since a single head in the Attention module is evaluated with one score but contains significantly more weights than a single channel in the MLP module, we apply scaling factors based on the model design to balance number of pruned parameters between the Attention heads and MLP channels. For a specified global pruning ratio ρ, hidden state dimension D, head dimension D_h in the Attention module, and intermediate size D_{inter} in the MLP module, we define Ψ as the set that stores the scores for the whole model, then apply the scaling factor α when generating the threshold η for all the scores as follows,

$$\Psi := \alpha \cdot (\{\{z_{h,l}^{\text{head}}\}_{h=1}^{H}\}_{l=1}^{L} \bigcup \{\{z_{i,l}^{\text{mlp}}\}_{i=1}^{D_{\text{inter}}}\}_{l=1}^{L}, \quad (8)$$

$$\eta = \text{sort}(\Psi)[(\rho(L \cdot H + L \cdot D_{\text{inter}}))],\ \alpha = \frac{4D_h}{3},$$

where $H = \text{index}(D/D_h)$ denotes the number of head in Attention module, L denotes the number of layers for the whole model. Since the Attention module involves pruning 4 linear projections (query, key, value, and output) in head level, while the MLP module prunes only 3 (up, gate, and down) in channel level, the scaling factor α is given by $\frac{4D_h}{3}$.

When the threshold η is determined, we prune the heads in the Attention module and the channels in the MLP module across all layers based on the strategy that removes heads or channels with numerical scores below the threshold.

3.4 Compensation for Pruning

With the above discussion, we obtain the pruning mask with Newton's method. To further improve the model performance, we modify the remaining weights in the model to compensate the loss of the pruned weights.

Problem Formulation. Note that to align the internal computations in the attention and MLP modules, we prune the rows of the output layers in the modules and the columns in other layers of the modules. If the columns of a layer with W is pruned in XW, the corresponding columns of the output also become zero and we are not able to compensate its loss, since modifying other unpruned columns can not change the zero output for the pruned columns. Thus, we only update the weights of the output layers with row pruning in the Attention and MLP modules. We modify the remaining rows based on pruned rows in W. For layers with column pruning, we do not modify their unpruned weights.

For the original weights W, after pruning, there are k pruned rows which are all zeros and their row indexes are denoted by $p_i, \forall i \in [k]$. We modify the weights with the weight perturbations δW, so that the layer output difference (before and after pruning) measured with ℓ_2 norm is minimized. The weight optimization problem can formulated as the following,

$$\begin{aligned} \min_{\delta W} \quad & \mathcal{L}(\delta W) = \|X(W+\delta W) - XW\|_2^2 = \|X\delta W\|_2^2, \\ \text{s.t.} \quad & e_{p_i}^\top \delta W + (W)_{p_i,*} = 0, \quad \text{for } i = 1, 2, \ldots, k. \end{aligned} \quad (9)$$

where $e_{p_i} \in \mathbb{R}^{D\times 1}$ is the one-hot vector with the p_i^{th} element as 1 and all others as 0. Thus, $e_{p_i}^\top \delta W$ denotes selecting the p_i^{th} row of δW. $(W)_{i,j}$ represents the element in the i^{th} row and j^{th} column of the matrix. Then, $(W)_{p_i,*}$ represents the p_i^{th} row of W. We can see that the constraint in Eq. (9) ensures that the corresponding pruned rows in the modified weights are all zeros, and the remaining weights are optimized to minimize the loss incurred by pruned rows.

It can be further transformed to the following,

$$\begin{aligned} \min_{\delta W} \quad & \mathcal{L}(\delta W) = \|X\delta W\|_2^2, \\ \text{s.t.} \quad & M_p^\top \delta W + W_p = 0, \end{aligned} \quad (10)$$

where $M_p \in \mathbb{R}^{D\times k}$ is the collection of all e_{p_i}, *i.e.*, $(M_p)_{*,i} = e_{p_i}$, or $(M_p^\top)_{i,*} = e_{p_i}^\top, \forall i \in [k]$. Similarly, W_p is a collection of all pruned rows in W with $(W_p)_{i,*} = (W)_{p_i,*}, \forall i \in [k]$. We have $W_p = M_p^\top W$.

Optimal Solution. Eq. (10) can be solved analytically with the following Theorem 3. The detailed proof is shown in Appendix B.

Theorem 3. *The optimal solution for Eq.* (10) *can be derived as the following,*

$$\delta W^* = -(2X^\top X)^{-1} M_p (M_p^\top (2X^\top X)^{-1} M_p)^{-1} M_p^\top W. \quad (11)$$

Remark 4. *The optimal loss of Problem* (10) *corresponding to the optimal weight perturbation can be expressed as*

$$L^* = \frac{1}{2} \sum_i (W^\top M_p (M_p^\top (2X^\top X)^{-1} M_p)^{-1} M_p^\top W)_{i,i}. \quad (12)$$

The sum in Eq. (12) *is computed over* D' *(the number of columns in* W*),* i.e.*,* $i \in [D']$.

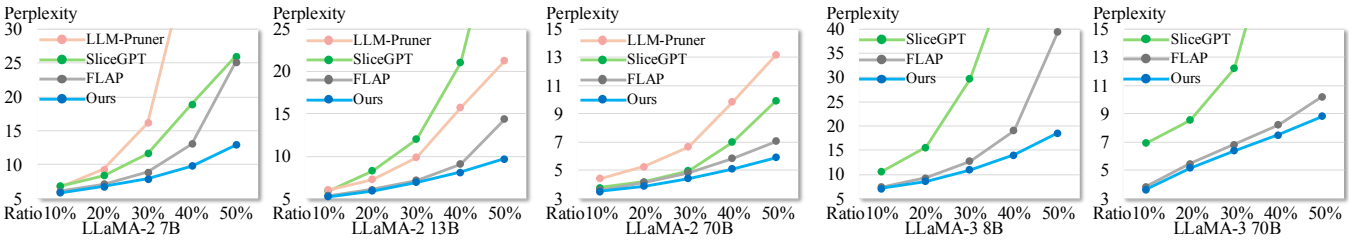

Figure 2: Perplexity (↓) results for LLaMA-2 and LLaMA-3 models on WikiText2 dataset with 2048 sequence length. Comprehensive detailed results are included in Table 6 and Table 7 at Appendix A.2 and A.3.

Method	Prune Ratio	LLaMA-7B Wiki	LLaMA-7B PTB	LLaMA-7B C4	LLaMA-13B Wiki	LLaMA-13B PTB	LLaMA-13B C4	LLaMA-30B Wiki	LLaMA-30B PTB	LLaMA-30B C4	LLaMA-65B Wiki	LLaMA-65B PTB	LLaMA-65B C4
Baseline	\	5.68	27.34	7.08	5.09	19.23	6.61	4.10	16.29	5.98	3.53	17.61	5.62
LLM-Pruner	10%	7.41	36.73	9.25	6.38	31.85	8.16	4.92	18.17	6.63	3.98	19.44	6.08
SliceGPT	10%	6.97	88.48	23.54	6.11	60.15	20.18	5.24	39.72	17.83	4.57	36.20	14.14
FLAP	10%	6.34	32.39	8.058	5.45	20.99	7.33	4.52	17.29	6.49	3.91	19.35	6.04
Ours	10%	**6.01**	**31.65**	**7.94**	**5.38**	**20.52**	**7.27**	**4.43**	**17.26**	**6.47**	**3.82**	**19.28**	**6.02**
LLM-Pruner	20%	10.73	59.73	12.15	6.38	31.85	9.42	5.83	20.18	7.55	4.65	21.85	6.75
SliceGPT	20%	8.42	120.89	35.93	7.17	86.26	29.70	6.18	50.95	26.85	5.34	61.09	21.86
FLAP	20%	7.40	36.77	9.99	6.03	23.33	8.42	5.18	19.30	7.42	4.45	21.45	6.75
Ours	20%	**6.60**	**35.75**	**9.49**	**5.89**	**23.11**	**8.39**	**4.92**	**18.58**	**7.36**	**4.26**	**20.94**	**6.73**
LLM-Pruner	30%	18.58	93.24	17.78	11.81	45.42	12.65	7.59	24.97	9.08	5.52	26.38	7.53
SliceGPT	30%	12.75	258.90	67.33	9.18	125.40	46.46	7.74	75.89	42.71	6.56	74.43	35.68
FLAP	30%	9.18	47.35	13.08	6.97	27.36	10.01	6.28	21.88	8.53	5.10	23.91	7.59
Ours	30%	**7.56**	**41.05**	**11.53**	**6.57**	**26.27**	**9.98**	**5.46**	**20.48**	**8.46**	**4.75**	**22.13**	**7.51**
LLM-Pruner	50%	126.0	460.7	73.88	45.69	152.99	36.94	19.68	78.29	18.64	9.34	43.79	12.16
SliceGPT	50%	1540	6364	4847	18.75	277.34	122.5	15.60	195.4	118.5	12.01	160.3	92.66
FLAP	50%	21.89	135.8	30.86	12.88	53.54	18.37	13.41	47.30	13.17	6.98	28.52	10.36
Ours	50%	**11.66**	**82.55**	**20.72**	**8.91**	**37.56**	**16.12**	**7.25**	**26.68**	**11.91**	**6.02**	**25.17**	**9.73**
LLM-Pruner	70%	9010	4111	2655	5900	6039	1334	895.7	3274	456.9	Nan	Nan	Nan
SliceGPT	70%	3605	7304	8096	67.65	874.9	537.4	71.25	633.1	406.6	102.4	863.9	662.8
FLAP	70%	577.9	1835	833.7	647.8	1588	975.1	2786	2735	2416	Nan	2333	Nan
Ours	70%	**162.9**	**721.3**	**361.6**	**41.66**	**275.7**	**115.3**	**39.88**	**124.2**	**50.43**	**9.65**	**69.49**	**20.84**

Table 1: Perplexity (↓) results for LLaMA-1 family models with different pruning ratios on WikiText2, PTB, and C4 with 2048 sequence length. Full results with larger sparsity ratios are included in Table 5 at Appendix A.1.

Remark 5. *If the rank of $2X^\top X$ is not full so that the inversion $(2X^\top X)^{-1}$ is unavailable, we apply the dampening method to compute $(2X^\top X + \gamma \cdot I)^{-1}$ instead of $(2X^\top X)^{-1}$, with γ as the dampening ratio.*

3.5 Complexity Analysis

For the computation of numerical score, according to the Lemma 1 and Theorem 2, the complexity is $O(TD^3)$ where T represents the number of iterations for Newton's Method, typically around 50 in practice. Additionally, for the compensation method, as demonstrated in Eq. (11), the complexity is $O(D^3)$ as we need to compute the inverse of a matrix. The matrix multiplication with M_p or M_p^T just selects the columns or rows of a matrix, without the need of actual multiplication. The complexity for numerical score calculation and compensation is the same with state-of-the-art methods, such as SparseGPT (Frantar and Alistarh 2023). In practice, the compensation is finished with just a few data samples on only the output projection layers of the Attention module and the down projection layers of the MLP module, which is more efficient compared with other recovery methods such as LLM-Pruner (Ma, Fang, and Wang 2023) to finetune the whole model on whole dataset, or SliceGPT (Ashkboos et al. 2024) to adopt a large amount of samples for calibration.

4 Experimental Results

4.1 Experiment Setup

We conduct the experiments on LLaMA model families including LLaMA-1 (Touvron et al. 2023a), LLaMA-2 (Touvron et al. 2023b), and LLaMA-3 (Meta 2024) for the language generation tasks. For evaluations, we compare the perplexity of the models on the WikiText2 (Merity et al. 2016), PTB (Marcus, Santorini, and Marcinkiewicz 1993), and C4 (Raffel et al. 2020) datasets with the 2048 sequence length. We also follow LLM-Pruner to evaluate the zero-shot accuracy on common sense reasoning zero-shot classification datasets including BoolQ (Clark et al. 2019), PIQA (Bisk et al. 2020), HellaSwag (Zellers et al. 2019),

Method	Prune Ratio	BoolQ	PIQA	Hella Swag	Wino Grande	ARC-e	ARC-c	OBQA	Average Acc.
LLaMA-7B	/	73.18	78.35	72.99	67.01	67.45	41.38	42.40	63.25
LLM-Pruner(v)	20%	61.44	71.71	57.27	54.22	55.77	33.96	38.40	53.25
LLM-Pruner(e2)		59.39	75.57	65.34	61.33	59.18	37.12	39.80	56.82
LLM-Pruner(e1)		57.06	75.68	66.80	59.83	60.94	36.52	40.00	56.69
SliceGPT	20%	37.89	64.09	45.67	62.75	53.62	31.74	33.20	46.99
FLAP	20%	68.59	74.21	64.98	64.40	59.89	37.80	40.20	58.58
Ours	20%	67.92	74.76	67.31	66.54	58.80	36.77	39.4	**58.79**

Table 2: Pruning results for LLaMA-7B on common sense reasoning datasets. LLM-Pruner (v) and (ei) denote vector-wise and element-wise with i-th order ($i = 1, 2$). Full results with more sparsity ratios and LLaMA-13B are in Table 9 at Appendix B.

Figure 3: Visualization of generated images through LlamaGen-3B in 384×384 resolution (cfg=1.65) with 10% sparsity.

WinoGrande (Sakaguchi et al. 2021), ARC-easy (Clark et al. 2018), ARC-challenge (Clark et al. 2018), and OpenbookQA (Mihaylov et al. 2018). For experiments, we adopt 128 samples from training dataset of WikiText2 to compute the numerical score and compensate the pruned models. For fairness, we also adopt 128 samples for other methods.

As for the image generation tasks, we adopt the LlamaGen (Sun et al. 2024) model family with LlamaGen-XXL and LlamaGen-3B to verify the effectiveness of our method on image generation tasks. We adopt the Fréchet inception distance (FID) (Heusel et al. 2017), Inception Score (IS) (Salimans et al. 2016), sFID (Nash et al. 2021), and Precision/Recall (Kynkäänniemi et al. 2019) as the evaluation metrics on ImageNet dataset (Deng et al. 2009). For all evaluations, we utilized ADM's TensorFlow scripts (Dhariwal and Nichol 2021) to ensure fair and consistent comparisons. Given that LLM-Pruner requires a backward process and SliceGPT has slow pruning, we further implement FLAP for comparative analysis in image generation tasks. In pratical, we generate 128 images for each class of ImageNet with LlamaGen models for the computation of numerical score and compensation. Same strategy for FLAP for fairness.

4.2 Results of LLMs

For the LLaMA models, we present the results with different pruning ratios varying from 10% to 70% in Table 1. Based on the perplexity results evaluated with 2048 sequence length on three datasets, our method consistently outperforms other methods across all pruning ratios, demonstrating the effectiveness of our proposed approach. Full results with more sparse ratios are included in Table 5 of Appendix A.1. Results show that for the larger model LLaMA-65B with pruning ratio of 70%, both LLM-Pruner and FLAP fail to produce an effective pruned model with their respective methods. In contrast, our method successfully maintains the most of the model's capabilities.

We further evaluate the zero-shot capabilities of the pruned model across seven downstream tasks. The results of LLaMA-7B model are shown in Table 2. Full results, including additional pruning ratios and the LLaMA-13B model, are detailed in Table 9 in Appendix A.5. Our method demonstrates superior performance compared to the other three methods on those common sense reasoning zero-shot classification datasets. Besides, we show the results with LLaMA and LLaMA-2 models of our method on MMLU (Hendrycks et al. 2021) and GSM8K (Cobbe et al. 2021) datasets in Table 8 of Appendix A.4, which demonstrates that our method retains both generative and mathematical capabilities.

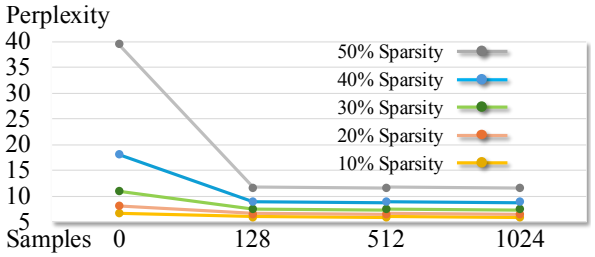

Figure 4: Ablation for number of samples for compensation.

We show the results for LLaMA-2 and LLaMA-3 models with 2048 sequence length on WikiText2 dataset in Figure 2. The detailed perplexity results for both model families on three datasets are shown in Table 6 and Table 7 of Appendix A.2 and A.3. The blue line representing our method's results consistently appears at the lowest position on the graphs, indicating its superior performance compared to the other methods with all model families.

4.3 Results of Image Generation

We implement the FLAP pruning method on LlamaGen model and compare this method on image generation task. We show the sparse results with LlamaGen-XXL (1.4B) and LlamaGen-3B models on ImageNet with 384×384 resolution in Table 3. We observe that for the smaller model LlamaGen-XXL (1.4B), our method shows a distinct advantage at higher pruning ratios. For the larger model LlamaGen-3B, our method consistently outperforms across all pruning ratios, effectively preserving most of the original model's capabilities. We further visualize the images generated by 10% sparsity models in Figure 3 with additional visualizations provided in Figure 6 of Appendix A.6. We observe that our method generates better image results compared to FLAP method in most cases.

4.4 Ablation Study

Results with 128 Sequence Length. To demonstrate the effectiveness of our method for short sequence lengths, we present the results generated with a sequence length of 128 in Table 4 using the LLaMA-7B model and the WikiText2 dataset. Comprehensive results, including additional pruning ratios and datasets, are provided in Table 10 of Appendix A.7. As observed, our method consistently performs the best across all pruning ratios.

Number of Samples for Compensation. To verify the efficiency of the compensation process for our method, we conducted experiments using different numbers of samples. The results of these experiments are shown in Figure 4. The results demonstrate that the performance difference between compensation with 128 samples versus 512 or even 1024 samples is minimal across all pruning ratios. This indicates that 128 samples are sufficient for our compensation method, highlighting its efficiency.

Memory & Generation Speed. We show the memory reduction and generation acceleration in Figure 5. The results

Method	Ratio	FID ↓	sFID ↓	IS ↑	Prec ↑	Rec ↑
LlamaGen-XXL (cfg=1.75)						
/	/	2.39	6.02	253.16	80.73%	59.60%
FLAP	10%	7.87	9.92	145.25	61.96%	63.34%
Ours	10%	**6.09**	**7.70**	**168.96**	**70.98%**	**65.01%**
FLAP	15%	15.93	11.81	100.48	52.05%	62.02%
Ours	15%	**11.29**	**9.85**	**124.31**	**62.95%**	**65.84%**
FLAP	20%	53.86	20.63	32.41	28.31%	67.14%
Ours	20%	**22.45**	**14.16**	**78.64**	**53.68%**	**67.66%**
LlamaGen-3B (cfg=1.65)						
/	/	2.26	6.19	260.46	82.07%	58.35%
FLAP	10%	7.57	8.40	158.74	67.19%	64.90%
Ours	10%	**3.97**	**7.45**	**202.93**	**73.93%**	62.86%
FLAP	15%	38.45	23.61	57.29	43.45%	63.27%
Ours	15%	**8.92**	**10.32**	**152.36**	**66.83%**	**63.97%**
FLAP	20%	162.15	93.98	5.97	13.36%	28.03%
Ours	20%	**20.16**	**16.29**	**95.05**	**54.87%**	**63.72%**

Table 3: Sparse results for image generation task with LlamaGen model family in 384×384 resolution.

Prune Ratio	10%	20%	30%	40%	50%
LLM-Pruner	15.37	19.09	30.64	52.28	122.8
SliceGPT	14.52	19.27	44.96	535.5	2241
FLAP	13.84	14.62	17.62	22.32	31.80
Ours	**13.31**	**14.47**	**16.40**	**19.04**	**23.32**

Table 4: Results for LLaMA-7B model on WikiText2 with 128 sequence length. Full results with more datasets are in Table 10 at Appendix B.

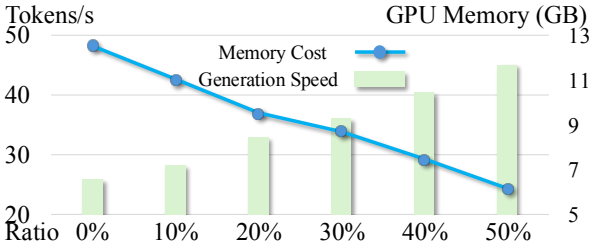

Figure 5: GPU memory v.s. generation speed.

are obtained using an NVIDIA A100 GPU with a sentence consisting of 64 tokens as the model input. The results show that as the pruning ratio increases, there is a corresponding decrease in GPU memory usage and an increase in generation speed, which validates the effectiveness of our method.

5 Conclusion and Limitation

In this paper, we propose numerical score through Newton's Method for the minimization of pruning errors. Also, we introduce a compensation algorithm to reconstruct weights. Results show that our method not only achieves the SOTA performance but also reduces memory usage and accelerates generation on GPUs. One limitation of our method is its reduced effectiveness for image generation tasks.

References

Achiam, J.; Adler, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

An, Y.; Zhao, X.; Yu, T.; Tang, M.; and Wang, J. 2023. Fluctuation-based Adaptive Structured Pruning for Large Language Models. arXiv:2312.11983.

Ashkboos, S.; et al. 2024. Slicegpt: Compress large language models by deleting rows and columns. *arXiv:2401.15024*.

Bisk, Y.; et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *AAAI*.

Bubeck, S.; et al. 2015. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4): 231–357.

Clark, C.; Lee, K.; Chang, M.-W.; Kwiatkowski, T.; Collins, M.; and Toutanova, K. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Dhariwal, P.; and Nichol, A. Q. 2021. Diffusion Models Beat GANs on Image Synthesis. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

Frantar, E.; and Alistarh, D. 2023. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *ICML*, 10323–10337. PMLR.

Gong, Y.; Zhan, Z.; Li, Z.; Niu, W.; Ma, X.; Wang, W.; Ren, B.; Ding, C.; Lin, X.; Xu, X.; et al. 2020. A privacy-preserving-oriented dnn pruning and mobile acceleration framework. In *Proceedings of the 2020 on Great Lakes Symposium on VLSI*, 119–124.

Gong, Y.; Zhan, Z.; Zhao, P.; et al. 2022. All-in-one: A highly representative dnn pruning framework for edge devices with dynamic power management. In *ICCAD*, 1–9.

Gong, Y.; Zhao, P.; Zhan, Z.; et al. 2023. Condense: A Framework for Device and Frequency Adaptive Neural Network Models on the Edge. In *DAC*, 1–6. IEEE.

Hassibi, B.; Stork, D.; and Wolff, G. 1993. Optimal Brain Surgeon and general network pruning. In *IEEE International Conference on Neural Networks*, 293–299 vol.1.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Jian, T.; Gong, Y.; Zhan, Z.; Shi, R.; Soltani, N.; Wang, Z.; Dy, J.; Chowdhury, K.; Wang, Y.; and Ioannidis, S. 2021. Radio frequency fingerprinting on the edge. *IEEE Transactions on Mobile Computing*, 21(11): 4078–4093.

Kong, Z.; Dong, P.; Ma, X.; Meng, X.; Niu, W.; Sun, M.; Shen, X.; Yuan, G.; Ren, B.; Tang, H.; et al. 2022. Spvit: Enabling faster vision transformers via latency-aware soft token pruning. In *ECCV*.

Kong, Z.; Ma, H.; Yuan, G.; Sun, M.; Xie, Y.; Dong, P.; Meng, X.; Shen, X.; Tang, H.; Qin, M.; et al. 2023. Peeling the onion: Hierarchical reduction of data redundancy for efficient vision transformer training. In *AAAI*.

Kynkäänniemi, T.; Karras, T.; Laine, S.; Lehtinen, J.; and Aila, T. 2019. Improved Precision and Recall Metric for Assessing Generative Models. *CoRR*, abs/1904.06991.

LeCun, Y.; Denker, J.; and Solla, S. 1989. Optimal Brain Damage. In Touretzky, D., ed., *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann.

Lee, D.; Kim, C.; Kim, S.; Cho, M.; and Han, W.-S. 2022. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11523–11532.

Li, T.; Tian, Y.; Li, H.; Deng, M.; and He, K. 2024a. Autoregressive Image Generation without Vector Quantization. *arXiv preprint arXiv:2406.11838*.

Li, Y.; Zhang, Y.; Liu, S.; and Lin, X. 2024b. Pruning then Reweighting: Towards Data-Efficient Training of Diffusion Models. *arXiv:2409.19128*.

Li, Y.; Zhao, P.; Ding, R.; Zhou, T.; Fei, Y.; Xu, X.; and Lin, X. 2024c. Neural architecture search for adversarial robustness via learnable pruning. *Frontiers in High Performance Computing*.

Li, Y.; Zhao, P.; Lin, X.; Kailkhura, B.; and Goldhahn, R. 2023. Less is More: Data Pruning for Faster Adversarial Training. *arXiv:2302.12366*.

Li, Y.; Zhao, P.; Yuan, G.; Lin, X.; Wang, Y.; and Chen, X. 2022. Pruning-as-search: Efficient neural architecture search via channel pruning and structural reparameterization. *arXiv preprint arXiv:2206.01198*.

Ma, X.; Fang, G.; and Wang, X. 2023. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36: 21702–21720.

Marcus, M. P.; Santorini, B.; and Marcinkiewicz, M. A. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 313–330.

Merity, S.; Xiong, C.; Bradbury, J.; and Socher, R. 2016. Pointer sentinel mixture models. *arXiv*.

Meta. 2024. LLaMA 3: The most capable openly available LLM to date. *https://ai.meta.com/blog/meta-llama-3/*.

Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *EMNLP*, 2381–2391. Brussels, Belgium: ACL.

Nash, C.; Menick, J.; Dieleman, S.; and Battaglia, P. W. 2021. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*.

Parmar, N.; et al. 2018. Image transformer. In *International conference on machine learning*, 4055–4064. PMLR.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative adversarial text to image synthesis. In *ICML*, 1060–1069. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.

Sakaguchi, K.; Bras, R. L.; Bhagavatula, C.; and Choi, Y. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9): 99–106.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. *NeurIPS*, 29.

Shen, X.; Dong, P.; Lu, L.; et al. 2024a. Agile-Quant: Activation-Guided Quantization for Faster Inference of LLMs on the Edge. In *AAAI*.

Shen, X.; Han, Z.; Lu, L.; et al. 2024b. HotaQ: Hardware Oriented Token Adaptive Quantization for Large Language Models. *TCAD*.

Shen, X.; Kong, Z.; Qin, M.; et al. 2023a. Data level lottery ticket hypothesis for vision transformers. *IJCAI*.

Shen, X.; Kong, Z.; Yang, C.; et al. 2024c. EdgeQAT: Entropy and Distribution Guided Quantization-Aware Training for the Acceleration of Lightweight LLMs on the Edge. *arXiv preprint arXiv:2402.10787*.

Shen, X.; Wang, Y.; Lin, M.; et al. 2023b. DeepMAD: Mathematical Architecture Design for Deep Convolutional Neural Network. In *CVPR*.

Shen, X.; Zhao, P.; Gong, Y.; Kong, Z.; Zhan, Z.; Wu, Y.; Lin, M.; Wu, C.; Lin, X.; and Wang, Y. 2024d. Search for Efficient Large Language Models. In *NeurIPS*.

Sun, M.; Liu, Z.; Bair, A.; and Kolter, J. Z. 2023. A Simple and Effective Pruning Approach for Large Language Models. *arXiv preprint arXiv:2306.11695*.

Sun, P.; Jiang, Y.; Chen, S.; Zhang, S.; Peng, B.; Luo, P.; and Yuan, Z. 2024. Autoregressive Model Beats Diffusion: Llama for Scalable Image Generation. *arXiv preprint arXiv:2406.06525*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wu, Y.; Gong, Y.; Zhao, P.; et al. 2022. Compiler-aware neural architecture search for on-mobile real-time super-resolution. In *ECCV*, 92–111. Springer.

Yang, Y.-Y.; Rashtchian, C.; Zhang, H.; Salakhutdinov, R. R.; and Chaudhuri, K. 2020. A Closer Look at Accuracy vs. Robustness. In *NeurIPS*, volume 33, 8588–8601. Curran Associates, Inc.

Yu, J.; Li, X.; Koh, J. Y.; Zhang, H.; Pang, R.; Qin, J.; Ku, A.; Xu, Y.; Baldridge, J.; and Wu, Y. 2021. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*.

Yu, J.; Xu, Y.; Koh, J. Y.; Luong, T.; Baid, G.; Wang, Z.; Vasudevan, V.; Ku, A.; Yang, Y.; Ayan, B. K.; et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3): 5.

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Zhan, Z.; Gong, Y.; Zhao, P.; Yuan, G.; et al. 2021. Achieving on-mobile real-time super-resolution with neural architecture and pruning search. In *ICCV*, 4821–4831.

Zhan, Z.; Kong, Z.; Gong, Y.; et al. 2024a. Exploring Token Pruning in Vision State Space Models. In *NeurIPS*.

Zhan, Z.; Wu, Y.; Gong, Y.; et al. 2024b. Fast and Memory-Efficient Video Diffusion Using Streamlined Inference. In *NeurIPS*.

Zhan, Z.; Wu, Y.; Kong, Z.; et al. 2024c. Rethinking Token Reduction for State Space Models. In *EMNLP*, 1686–1697. Miami, Florida, USA: ACL.

Zhang, Y.; Yao, Y.; Ram, P.; Zhao, P.; Chen, T.; Hong, M.; Wang, Y.; and Liu, S. 2022. Advancing model pruning via bi-level optimization. *Advances in Neural Information Processing Systems*, 35: 18309–18326.

Zhao, P.; Sun, F.; Shen, X.; et al. 2024. Pruning Foundation Models for High Accuracy without Retraining. In *Findings of EMNLP 2024*, 9681–9694. ACL.