Attention Weight Decomposition for Vision Model Compression

Hyunwoo Yu [*1] **Yubin Cho** [*1] **Kyeongbo Kong** [†2] **Suk-Ju Kang** [†1]

Abstract

Weight decomposition is a practical compression approach, yet prior methods overlook that multi-head attention consists of paired linear projections (Q–K and V–O). We propose *Uni*lateral decomposition for attention *Com*pression (*UniCom*), which decomposes only one side of each pair to better manage approximation error while leveraging attention's linear operations. Since Q-K-V-O weights exhibit different low-rank sensitivities across heads, UniCom adaptively chooses which side to decompose based on rank sensitivity. Experiments show strong generalization across vision and vision–language tasks. Under 60% MHA reduction, UniCom improves Top-1 by +11.2% / +5.7% / +6.4% over combined decomposition on DeiT-S / DeiT-S(distill) / DeiT-B, respectively.

1. Introduction

Transformers (Vaswani et al., 2017; Dosovitskiy et al., 2020; Touvron et al., 2021; Radford et al., 2021; Touvron et al., 2022; Liu et al., 2021; Wang et al., 2021) achieve high performance but suffer from substantial computational costs. Among various compression techniques like quantization and pruning, we focus on weight decomposition to enable efficient inference without fine-tuning.

For weight decomposition (Wang et al., 2025a; Yang et al., 2024; Saha et al., 2023; Qin et al., 2025; Pletenev et al., 2023), low-rank factorization with Singular Value Decomposition (SVD) offers an promising approach for compressing transformer-based models (Wu et al., 2023; Yu & Wu, 2023; Ashkboos et al., 2024; Chang et al., 2025). Recent studies (Tian et al., 2025; Chen et al., 2021; Yuan et al., 2023; Wang et al., 2024; 2025b; Thoma et al., 2025) have investigated activation-guided weight importance to enhance decomposition quality. However, existing per-weight decomposition exhibit an inefficiency in attention where each Query (Q), Key (K), Value (V), and Output (O) weights is factorized into two low-rank matrices and used as two sub-layers. As a result, substantial efficiency gains can only be realized when the reduced rank is smaller than $2/d$. However, this inherent inefficiency can be effectively overcome by explicitly exploiting the linear characteristic of the attention mechanism in vision models, where the Q-K and V-O computations involve no intermediate non-linear functions.

In this paper, we propose *Uni*lateral singular value decomposition method for vision model *Com*pression (*UniCom*), a novel and effective method that applies decomposition to only one side of the Q-K and V-O weights, thereby addressing the inefficiency of previous per-weight decomposition methods. When decomposing unilaterally, we focus on that each weight within the Q-K and V-O pairs exhibits different sensitivities to low-rank approximation across heads. Therefore, our UniCom dynamically selects which weight matrix to decompose based on the approximation error. After decomposing one side of the weight matrices, we multiply it with the other side, thereby reducing the overall weight while preserving the information of the non-decomposed side, and still achieving the effect of decomposing both sides.

Our experiments demonstrate that UniCom effectively preserves performance across diverse attention-based vision models and generalizes well to various downstream vision and vision-language tasks. Furthermore, it achieves a favorable trade-off among FLOPs, parameter reduction, and accuracy, particularly under challenging low-rank settings.

Our contributions are summarized as follows:

- We propose Unilateral singular value decomposition for vision model Compression (*UniCom*), a novel method that decomposes only one side of Q-K and V-O weights, effectively mitigating performance degradation commonly observed in low-rank approximations.
- UniCom leverages a dynamic selection strategy, where the decomposed side is determined head-wise based on low-rank approximation error.

[1]Department of Electronics Engineering, Sogang University, South Korea [2]Department of Electrical & Electronics Engineering, Pusan National University, South Korea. Correspondence to: Kyeongbo Kong <kbkong@pusan.ac.kr>, Suk-Ju Kang <sjkang@sogang.ac.kr>.

Workshop on Weight-Space Symmetries, held in conjunction with the 43rd International Conference on Machine Learning, Seoul, South Korea. 2026. Copyright 2026 by the author(s).

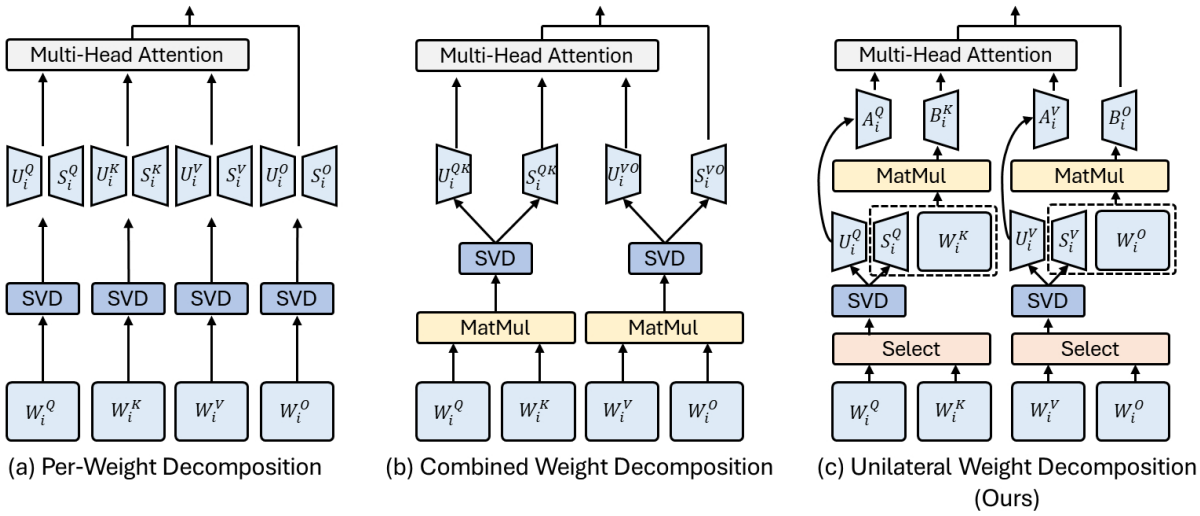

Figure 1. Comparison of different weight decomposition methods in multi-head attention: (a) Per-weight decomposition, (b) Combined weight decomposition, and (c) Unilateral weight decomposition (UniCom). UniCom illustrates the case where W^Q and W^V are selected.

- We show that UniCom generalizes across vision classification and vision–language tasks while consistently preserving model performance more effectively.

2. Method

2.1. Reformulation of Multi-Head Attention

Our weight decomposition method leverages the linear transformation property in the Q-K and V-O operations of MHA, a characteristic that clearly differentiates it from feedforward layers. In the MHA mechanism, given an input $X \in \mathbb{R}^{N\times C}$, the per-head projections Q_i, K_i, and V_i are generated through their respective weight matrices W_i^Q, W_i^K, and $W_i^V \in \mathbb{R}^{C\times d_h}$. The head-wise scaled dot-product attention is computed to obtain head_i, and finally, the concatenated outputs from all heads are passed through the projection matrix $W^O \in \mathbb{R}^{C\times C}$, yielding the final output representation. This can be formulated as follows:

$$\begin{aligned}\mathrm{head}_i &= \mathrm{Attention}(XW_i^Q, XW_i^K, XW_i^V)\\ &= \mathrm{Softmax}\left(\frac{XW_i^Q(XW_i^K)^\top}{\sqrt{d_h}}\right)XW_i^V,\end{aligned} \tag{1}$$

$$\mathrm{MHA}(X) = \mathrm{Concat}(\mathrm{head_1}, \ldots, \mathrm{head_h})W^O, \tag{2}$$

where Concat denotes the concatenation operator along the head dimension, d_h represents the dimension of each head, and h is the total number of heads. To explicitly describe the head-wise operations among W_i^Q, W_i^K, W_i^V, and W^O, we reformulate MHA such that the output projection W^O is divided across heads, with each head associated with $W_i^O \in \mathbb{R}^{d_h\times C}$, as follows:

$$\begin{aligned}\mathrm{MHA}(X) &= \sum_{i=1}^{h}\mathrm{head_i}W_i^O\\ &= \sum_{i=1}^{h}\mathrm{Softmax}\left(\frac{X(W_i^Q{W_i^K}^\top)X^\top}{\sqrt{d_h}}\right)X(W_i^V W_i^O).\end{aligned} \tag{3}$$

Therefore, MHA can be reformulated as a series of matrix multiplications between W_i^Q-W_i^K and W_i^V-W_i^O for the linear transformations of each head.

2.2. Conventional SVD for Attention

Based on the formulation of MHA in Eq. (3), the conventional per-weight decomposition method performs low-rank approximation by applying SVD to each of the projection matrices W_i^Q, W_i^K, W_i^V, and W_i^O. The SVD is employed to decompose the weight matrix $W \in \mathbb{R}^{in\times out}$ into a low-rank factorization of rank r, expressed as $W \approx U\Sigma S'$, where $U \in \mathbb{R}^{in\times r}$, $S' \in \mathbb{R}^{r\times out}$, and $\Sigma \in \mathbb{R}^{r\times r}$ is a diagonal matrix containing the singular values. Subsequently, Σ and S' are multiplied to form $S = \Sigma S'$, allowing the decomposition to be represented using two matrices, U and S, which are then used for the low-rank approximation. As shown in Figure 1 (a), performing the naive SVD for low-rank approximation on each weight matrix in MHA can be formulated as follows:

$$\mathrm{MHA}(X) \approx \sum_{i=1}^{h}\mathrm{Softmax}\left(\frac{X(U_i^Q S_i^Q {S_i^K}^\top {U_i^K}^\top)X^\top}{\sqrt{d_h}}\right)X(U_i^V S_i^V U_i^O S_i^O), \tag{4}$$

where $U_i^Q, U_i^K, U_i^V \in \mathbb{R}^{C\times r}$ and $U_i^O \in \mathbb{R}^{d_h\times r}$ denote the left singular vector matrices obtained from SVD, $S_i^Q, S_i^K, S_i^V \in \mathbb{R}^{r\times d_h}$ and $S_i^O \in \mathbb{R}^{r\times C}$ are the corresponding low-rank factor matrices. As shown in Figure 1 (b), combined decomposition method (Xiao et al., 2023) first performs matrix multiplication between W_i^Q-W_i^K and W_i^V-W_i^O, and then applies SVD to effectively allocate the decomposed weights. The low-rank factorized MHA can be formulated as follows:

$$\text{MHA}(X) \approx \sum_{i=1}^{h} \text{Softmax}\left(\frac{X(U_i^{QK}S_i^{QK^\top})X^\top}{\sqrt{r}}\right) X(U_i^{VO}S_i^{VO}), \tag{5}$$

where U_i^{QK} and $S_i^{QK} \in \mathbb{R}^{C\times r}$ denote low-rank factors obtained from the head-wise SVD of $W_i^Q W_i^{K^T}$. $U_i^{VO} \in \mathbb{R}^{C\times r}$ and $S_i^{VO} \in \mathbb{R}^{r\times C}$ represent low-rank factors obtained from the head-wise SVD of $W_i^V W_i^O$. These low-rank factors, $U_i^{QK}, S_i^{QK}, U_i^{VO}$ and S_i^{VO}, are used as compressed projection weights for Q, K, V and O, respectively.

2.3. The Proposed Unilateral Weight Decomposition

Our UniCom selectively decomposes only one side of the weight pairs, W_i^Q-W_i^K and W_i^V-W_i^O, while achieving a more effective approximation than applying low-rank factorization to both sides. Since the rank sensitivity of these pairs varies across heads, we first evaluate the relative sensitivity by computing a Frobenius norm-based loss. This measure allows us to identify the less rank-sensitive weight within each pair, thereby guiding the selection of the side to be decomposed. The Frobenius norm-based loss is defined as:

$$\begin{aligned} \text{SVD}(W) &= \tilde{W}(r) = US, \\ L_I(W, r) &= ||W - \tilde{W}(r)||_F, \end{aligned} \tag{6}$$

where $\tilde{W}(r)$ denotes the rank r approximation of W, obtained via truncated SVD. This formulation provides a principled criterion for selecting the decomposition side. Based on this criterion, we determine which weight to be decomposed within each pair, either W_i^Q or W_i^K, and W_i^V or W_i^O. After that, we apply SVD only to the selected weight as shown in Figure 1 (c). The resulting low-rank factorization is then combined with its non-decomposed counterpart via a linear operator. This enables the model to preserve information from the intact side while reducing the computational complexity on the decomposed side. This process is formulated as follows:

$$\begin{aligned} \delta_i^{\text{QK}} &= 1\left[\ell(W_i^Q, r_{QK}) \le \ell(W_i^K, r_{QK})\right], \\ \beta_i^{\text{VO}} &= 1\left[\ell(W_i^V, r_{VO}) \le \ell(W_i^O, r_{VO})\right], \end{aligned} \tag{7}$$

$$\begin{aligned} A_i^Q &= \delta_i^{\text{QK}} U_i^Q + (1-\delta_i^{\text{QK}}) W_i^Q S_i^{K^\top}, \\ B_i^K &= \delta_i^{\text{QK}} S_i^Q W_i^{K^\top} + (1-\delta_i^{\text{QK}}) U_i^{K^\top}, \end{aligned} \tag{8}$$

$$\begin{aligned} A_i^V &= \beta_i^{\text{VO}} U_i^V + (1-\beta_i^{\text{VO}}) W_i^V S_i^{O^\top}, \\ B_i^O &= \beta_i^{\text{VO}} S_i^V W_i^{O^\top} + (1-\beta_i^{\text{VO}}) U_i^{O^\top}, \end{aligned} \tag{9}$$

where $A_i^Q \in \mathbb{R}^{C\times r_{QK}}$, $B_i^K \in \mathbb{R}^{r_{QK}\times C}$, $A_i^V \in \mathbb{R}^{C\times r_{VO}}$, and $B_i^O \in \mathbb{R}^{r_{VO}\times C}$ denote the selectively compressed weights. These matrices are used as the effective projection weights for Q, K, V, and O, respectively. r_{QK} and r_{VO} represent the ranks for low-rank approximation. In practice, we observed that the rank sensitivity of the W_i^Q-W_i^K pair differs from that of the W_i^V-W_i^O pair. Therefore, we assign different target ranks for each pair and further explore a simple hierarchical strategy that adapts the rank settings across layers. Given a predefined pair (r_{min},r_{max}), the ranks are assigned layer-wise according to an arithmetic progression between these two values. Accordingly, the low-rank factorized MHA with UniCom is formulated as:

$$\text{MHA}(X) \approx \sum_{i=1}^{h} \text{Softmax}\left(\frac{X(A_i^Q B_i^K)X^\top}{\sqrt{r_{QK}}}\right) X(A_i^V B_i^O). \tag{10}$$

This formulation ensures that the model retains information on the intact side while achieving the effect of decomposing both sides. Compared to the naive SVD and the combined SVD approaches, our UniCom leads to a better performance-efficiency trade-off with respect to the number of parameters and FLOPs due to a more effective decomposition strategy.

3. Experiments

3.1. UniCom for Large Vision Models

To evaluate the effectiveness of our method, we compare our UniCom with existing per-weight and combined weight decomposition approaches on various large-scale vision models on ImageNet-1K. In Table 1, UniCom consistently achieves higher accuracy than two different decomposition strategies under comparable parameter budgets in Multi-Head Attention (MHA). In particular, while all decomposition methods reduce the number of MHA parameters and inference latency, our unilateral decomposition preserves performance more effectively than per-weight and combined weight decomposition methods across diverse model architectures. These results demonstrate that UniCom generalizes well to different backbone designs and scales, supporting its scalability and extensibility to modern large transformer-based vision models.

3.2. UniCom for Large Multi-modal Model

To further validate the effectiveness of UniCom in comparison with the combined weight decomposition method, we conduct extensive experiments on vision-language models, including EVA-CLIP and LLaVA1.5. For the retrieval tasks, we report both image-to-text (I2T) and text-to-image (T2I) retrieval performance using Recall@1 (R@1). For

Table 1. Comparison with the existing weight decomposition methods and our UniCom for various vision models on ImageNet-1K.

Method	DINOv2 (Oquab et al., 2023)			DeiT3-Large (Touvron et al., 2022)		
	Params in MHA	Top-1 ↑	Latency ↓	Params in MHA	Top-1 ↑	Latency ↓
Baseline	28.3	81.2	38.6	100.7	86.8	99.1
Per-Weight Decomposition	22.5	65.5	37.3	50.3	82.3	91.7
Combined Weight Decomposition	22.5	64.2	36.0	50.3	81.7	84.5
Unilateral Weight Decomposition (Ours)	22.5	**73.2**	35.7	50.3	**83.0**	81.9
Method	ViT-Large (Dosovitskiy et al., 2020)			Swin-Large (Liu et al., 2021)		
	Params in MHA	Top-1 ↑	Latency ↓	Params in MHA	Top-1 ↑	Latency ↓
Baseline	100.7	84.3	100.6	62.8	86.3	67.8
Per-Weight Decomposition	50.3	75.0	91.4	31.4	80.4	63.9
Combined Weight Decomposition	50.3	74.9	83.6	31.4	79.2	58.3
Unilateral Weight Decomposition (Ours)	50.3	**77.7**	81.1	31.4	**81.2**	57.3

Table 2. Comparison of the combined weight decomposition method and our UniCom on vision-language models. EVA-CLIP is evaluated on Flickr30k. LLaVA1.5 is evaluated on widely used VQA (*i.e.*, VQAv2 and GQA) and multimodal evaluation (*i.e.*, MME) benchmarks.

3*Method	EVA-CLIP (Sun et al., 2023)				LLaVA1.5-7B (Liu et al., 2024)				
	Params (M)	GFLOPs	I2T Retrieval	T2I Retrieval	Params (M) of Vision Tower	GFLOPs	VQAv2	GQA	MME
Baseline	1136.6	299.9	90.0	74.6	303.5	191.1	78.5	61.9	1504.6
Combined Weight Decomposition	1021.2	267.6	88.3	73.7	272.0	167.8	73.6	59.4	1374.7
Unilateral Weight Decomposition (Ours)	1021.2	267.6	**89.3**	**74.0**	272.0	167.8	**74.7**	**59.8**	**1408.0**

Table 3. Comparison with the per-weight and the combined weight decomposition methods on different parameter reduction ratios.

Method	DeiT-Small (79.9%)			DeiT-S distill (80.9%)			DeiT-Base (81.8%)		
# of Params. in MHA ↓	20%	40%	60%	20%	40%	60%	20%	40%	60%
Per-Weight SVD	75.7	67.2	31.4	73.0	56.2	23.0	80.7	79.0	71.4
Combined SVD	78.3	74.2	54.2	79.8	76.1	63.6	81.6	79.4	65.7
UniCom (Ours)	**78.4**	**74.4**	**65.4**	**79.9**	**76.5**	**69.3**	**81.7**	**80.4**	**72.1**

LLaVA-1.5-7B, since both the proposed method and the combined method target the vision attention modules, we apply them to the vision tower. As shown in Table 2, UniCom achieves superior performance compared to the combined weight decomposition method under similar parameter and GFLOPs budgets. In particular, despite comparable reductions in parameters and computational cost, UniCom improves I2T retrieval performance by 1.0% and VQAv2 performance by 1.1% over the combined method. Although the performance gains on T2I retrieval and GQA are relatively modest, UniCom consistently matches or outperforms the combined method across a range of multi-modal and reasoning-oriented tasks. These results show that UniCom preserves task-relevant representations and generalizes well to diverse vision–language models.

3.3. Comparison on Different Reduction Ratios

As shown in Table 3, we conduct experiments by varying the ranks in the MHA layers to set different levels of the parameter reduction ratio. The results consistently demonstrate that our unilateral decomposition method achieves higher performance across all parameter reduction settings. In particular, under a large reduction ratio of 60%, our method significantly outperforms the combined weight decomposition with improvements of 11.2%, 5.7%, and 6.4% for DeiT-Small, DeiT-Small distilled and DeiT-Base, respectively. Moreover, for DeiT-Base, even when the parameter count is reduced by 20%, the resulting accuracy drop remains extremely small, only about 0.1%, which highlights the stability of our approach under light compression as well. These results indicate that decomposing only unilateral weight pairs (W_i^Q-W_i^K, W_i^V-W_i^O) effectively mitigates information loss from the original weights, thus making our method a robust and efficient decomposition strategy for attention-based vision models compared to two existing decomposition methods.

4. Conclusion

We proposed a novel singular value decomposition method for vision model compression, *UniCom*, which applies low-rank approximation by decomposing only one side of each weight pair (W^Q-W^K and W^V-W^O) in the multi-head attention of Transformer-based vision models. UniCom dynamically selected the weights to decompose based on their sensitivity to low-rank approximation across heads, thereby better preserving information contained in rank-sensitive weights. Our approach overcomes the inefficiency of per-weight decomposition methods and achieves better performance retention than combined decomposition methods, especially in low-rank approximation. Experimental results demonstrated that our method effectively minimizes information loss compared to per-weight and

combined decomposition methods. Furthermore, we show that UniCom generalizes across vision classification and vision-language tasks while consistently preserving model performance more effectively.

Acknowledgements

This research was supported by the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ITRC (Information Technology Research Center) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2026-RS-2023-00260091, 50%). In addition, this work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2025-02263706, Development of an analog-digital mixed ultra-low power neuromorphic edge SoC, 50%).

References

Ashkboos, S., Croci, M. L., Nascimento, M. G. d., Hoefler, T., and Hensman, J. Slicegpt: Compress large language models by deleting rows and columns. *arXiv preprint arXiv:2401.15024*, 2024.

Chang, C.-C., Lin, C.-Y., Akhauri, Y., Lin, W.-C., Wu, K.-C., Ceze, L., and Abdelfattah, M. S. xkv: Cross-layer svd for kv-cache compression. *arXiv preprint arXiv:2503.18893*, 2025.

Chen, P., Yu, H.-F., Dhillon, I., and Hsieh, C.-J. Drone: Data-aware low-rank compression for large nlp models. *Advances in neural information processing systems*, 34: 29321–29334, 2021.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Pletenev, S., Chekalina, V., Moskovskiy, D., Seleznev, M., Zagoruyko, S., and Panchenko, A. A computational study of matrix decomposition methods for compression of pre-trained transformers. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pp. 723–742, 2023.

Qin, B., Li, J., Tang, S., and Zhuang, Y. Dba: Efficient transformer with dynamic bilinear low-rank attention. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Saha, R., Srivastava, V., and Pilanci, M. Matrix compression via randomized low rank and low precision factorization. *Advances in Neural Information Processing Systems*, 36: 18828–18872, 2023.

Sun, Q., Fang, Y., Wu, L., Wang, X., and Cao, Y. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Thoma, M., Villasante, J., Aghajanzadeh, E., Sampath, S. B., Mori, P., Groetzinger, M., Dylkin, D., Vemparala, M.-R., Fasfous, N., Frickenstein, A., et al. Flar-svd: Fast and latency-aware singular value decomposition for model compression. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1898–1907, 2025.

Tian, J., Solgi, R., Lu, J., Yang, Y., Li, H., and Zhang, Z. Flat-llm: Fine-grained low-rank activation space transformation for large language model compression. *arXiv preprint arXiv:2505.23966*, 2025.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Touvron, H., Cord, M., and Jégou, H. Deit iii: Revenge of the vit. In *European conference on computer vision*, pp. 516–533. Springer, 2022.

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

Wang, Q., Ke, J., Tomizuka, M., Chen, Y., Keutzer, K., and Xu, C. Dobi-svd: Differentiable svd for llm compression and some new perspectives. *arXiv preprint arXiv:2502.02723*, 2025a.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., and Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 568–578, 2021.

Wang, X., Zheng, Y., Wan, Z., and Zhang, M. Svd-llm: Truncation-aware singular value decomposition for large language model compression. *arXiv preprint arXiv:2403.07378*, 2024.

Wang, X., Alam, S., Wan, Z., Shen, H., and Zhang, M. Svd-llm v2: Optimizing singular value truncation for large language model compression. *arXiv preprint arXiv:2503.12340*, 2025b.

Wu, Y., Kan, S., Zeng, M., and Li, M. Singularformer: Learning to decompose self-attention to linearize the complexity of transformer. In *IJCAI*, pp. 4433–4441, 2023.

Xiao, J., Yin, M., Gong, Y., Zang, X., Ren, J., and Yuan, B. Comcat: towards efficient compression and customization of attention-based vision models. *arXiv preprint arXiv:2305.17235*, 2023.

Yang, C., Sui, Y., Xiao, J., Huang, L., Gong, Y., Duan, Y., Jia, W., Yin, M., Cheng, Y., and Yuan, B. Moe-i2: Compressing mixture of experts models through inter-expert pruning and intra-expert low-rank decomposition. *arXiv preprint arXiv:2411.01016*, 2024.

Yu, H. and Wu, J. Compressing transformers: features are low-rank, but weights are not! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11007–11015, 2023.

Yuan, Z., Shang, Y., Song, Y., Wu, Q., Yan, Y., and Sun, G. Asvd: Activation-aware singular value decomposition for compressing large language models. *arXiv preprint arXiv:2312.05821*, 2023.