EXPLOITING WEIGHT SPACE SYMMETRIES FOR VISION ATTENTION COMPRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformers have achieved remarkable success across diverse domains, yet their large model sizes incur considerable computation and latency. Accordingly, weight decomposition has emerged as an effective compression technique. However, previous methods do not consider the unique mechanism of Multi-Head Attention (MHA), where query-key (Q-K) and value–output (V-O) computations are linear operations. To effectively leverage this mechanism, we propose **Uni**lateral **Dec**omposition for vision model compression (**UniDec**), a novel framework that applies decomposition to only one side of the Q-K or V-O weight pairs. This approach effectively addresses approximation error at the unilateral weight level while exploiting the linear characteristics of the attention. Additionally, since Q-K-V-O weights exhibit varying sensitivities to low-rank approximation across heads, UniDec adaptively selects which side to be decomposed according to the rank sensitivity, thereby preserving the important information of weights. Extensive experiments demonstrate that UniDec generalizes to various transformer-based structures. Demonstrating exceptional performance even under aggressive compression (*i.e.*, 60% MHA reduction), UniDec improves Top-1 accuracy by +34.0% / +46.3% on DeiT-Small / DeiT-S Distill while maintaining a +0.7% gain on DeiT-Base.

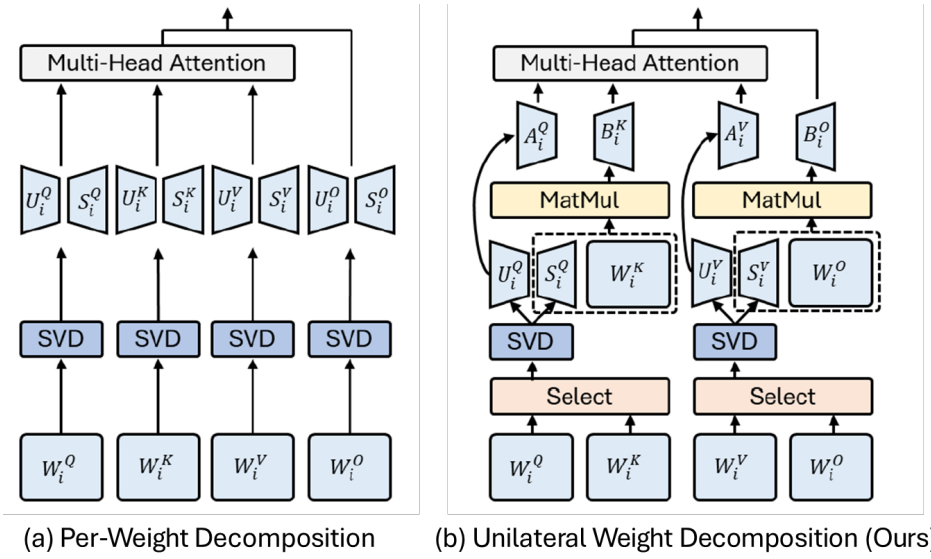
1 INTRODUCTION

Transformers (Dosovitskiy et al., 2020; Touvron et al., 2021; Radford et al., 2021; Touvron et al., 2022; Liu et al., 2021; Wang et al., 2021) have achieved remarkable success across diverse tasks by modeling long-range dependencies with attention. However, their increasing scale leads to substantial computational and latency costs. To mitigate these issues, various model compression techniques have been studied, including quantization, distillation, pruning, and weight decomposition. In this work, we focus on weight decomposition methods that enable efficient inference without fine-tuning.

For weight decomposition (Wang et al., 2025a; Yang et al., 2024a; 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) have investigated activation-guided weight importance to enhance decomposition quality. For example, ASVD (Yuan et al., 2023) introduces a diagonal scaling matrix to mitigate the distribution shift in activations before and after truncation. SVD-LLM (Wang et al., 2024; 2025b) apply a truncation-aware whitening strategy that attains the theoretical minimum loss. To enhance stability, FLAR-SVD (Thoma et al., 2025) integrates a Ledoit-Wolf shrinkage estimator into Cholesky SVD framework for vision models. 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 $2/d$.

When the weight decomposition is applied to attention-based vision models, these methods do not consider the fundamental linear characteristic of Multi-Head Attention (MHA). A notable characteristic of the MHA is that the Q-K and V-O computations involve no intermediate non-linear functions. Despite this linear coupling, conventional per-weight decomposition independently factorizes each

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.



(a) Per-Weight Decomposition (b) Unilateral Weight Decomposition (Ours)

Figure 1: Comparison of different weight decomposition methods in multi-head attention: (a) Per-weight decomposition and (b) Unilateral weight decomposition (UniDec). UniDec is illustrated for the case where W^Q and W^V are selected.

of the four weight matrices without exploiting the paired structure, causing approximation errors to accumulate across both sides and limiting the achievable efficiency gains.

In this paper, we propose **Uni**lateral singular value **Dec**omposition for vision model compression (**UniDec**), a novel and effective method that applies decomposition to only one side of the Q-K and V-O weights, thereby directly exploiting the linear coupling within each pair while preserving the full information of the non-decomposed counterpart. Unlike per-weight decomposition, which independently approximates each matrix and accumulates errors on both sides, UniDec concentrates the approximation on the less rank-sensitive side, achieving a more effective trade-off between compression and accuracy.Additionally, when decomposing unilaterally, we focus on the fact that each weight within the Q-K and V-O pairs exhibits different sensitivities to low-rank approximation across heads. Therefore, our UniDec 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. This approach reduces the overall weight while preserving the information of the non-decomposed side, ultimately achieving the effect of decomposing both sides.

Our experimental results show that UniDec preserves model performance more effectively across attention-based vision models with diverse scales and architectures. Moreover, UniDec demonstrates strong generalization capabilities across various prominent vision models, including ViT, DeiT, Swin, and DINO, consistently maintaining high performance. In addition, the results indicate that UniDec provides a favorable trade-off among FLOPs, parameter reduction, and model accuracy, particularly under more challenging low-rank approximation settings.

Our contributions are summarized as follows:

- We propose **Uni**lateral singular value **Dec**omposition for vision model compression (**UniDec**), 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.
- UniDec leverages a dynamic selection strategy, where the decomposed side is determined head-wise based on low-rank approximation error.
- Extensive experiments demonstrate that UniDec generalizes well across various transformer-based vision architectures, consistently and effectively preserving model performance.

2 Related Works

2.1 Model Compression for Transformers

Transformers achieve strong performance across NLP, vision, and multimodal tasks, but at the cost of high computational and latency. Various model compression techniques have been extensively explored to reduce the computations and accelerate the models. Knowledge distillation (Fang et al., 2021; Zhao et al., 2023) transfers behaviors from the complex teacher model to the compact student model via the softened logits, intermediate representation matching, and attention alignments. Pruning methods (Yang et al., 2023; Gao et al., 2024) set the less important weights to zero or remove the redundant components, such as weights, neurons or tokens. Quantization methods (Shang et al., 2023; Lv et al., 2024) compress the numerical model parameters from high-precision floating-point numbers to lower-precision integers. Recently, low-rank matrix decomposition (Li et al., 2023; Chang et al., 2024a) has been actively explored for model compression, which approximates high-dimensional weight matrices with lower-rank matrices. Our study focuses on training-free attention weight decomposition techniques for efficient model compression.

2.2 Weight Decomposition for NLP

Weight Decomposition (Chein et al., 2024; Yang et al., 2024b; Ji et al., 2024; Huang et al., 2025) has recently emerged as a promising approach for compressing and accelerating Transformers (Lee et al., 2024; Li et al., 2022). FWSVD (Hsu et al., 2022) introduced Fisher-weighted decomposition that leverages curvature information to preserve critical directions during truncation. ASVD (Yuan et al., 2023) proposed a diagonal scaling matrix, which scales the weight matrix based on the distribution patterns of input activation channels. SVD-LLM (Wang et al., 2024; 2025b) introduced a truncation-aware data whitening scheme to address the misalignment between the singular value truncation and the compression loss. ModeGPT (Lin et al., 2024) proposed a modular decomposition that jointly decomposes multiple matrices within each module by using SVD, Nyström and CR decomposition.

2.3 Weight Decomposition for Vision Models

The weight decomposition has also been recently studied for vision models (Veeramacheneni et al., 2025; Chang et al., 2024b; Jie & Deng, 2023). PELA (Guo et al., 2024) exploited low-rank approximation to entirely replace the pre-trained weights with the reduced low-rank matrices for downstream fine-tuning tasks. (Azizi et al., 2024) proposed an activation-aware mixed-rank compression strategy that adaptively allocates low-rank factors. Several studies (Thoma et al., 2025; Xiao et al., 2023) explored the weight decomposition without fine-tuning for vision models. FLAR-SVD (Thoma et al., 2025) used the Ledoit-Wolf shrinkage estimator for a stabilized Cholesky SVD. Our study exploits the property of attention in vision models but introduces a novel approach that unilaterally selects the weight pairs.

3 Method

Since our approach leverages the linear property of the Query(Q)-Key(K) and Value(V)-Output(O) operations within Multi-Head Attention (MHA), we first provide a brief overview of the MHA mechanism and then reformulate it in a head-wise representation. Then, we discuss existing decomposition methods that exploit this property. Finally, we provide a detailed description of the proposed unilateral singular value decomposition (UniDec), including its head-wise dynamic selection strategy.

3.1 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 $W_i^V \in \mathbb{R}^{C\times d_h}$. Following this step, 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:

$$\text{head}_i = \text{Attention}(XW_i^Q, XW_i^K, XW_i^V) = \text{Softmax}\left(\frac{XW_i^Q(XW_i^K)^\top}{\sqrt{d_h}}\right)XW_i^V. \quad (1)$$

$$\text{MHA}(X) = \text{Concat}(\text{head}_1, \ldots, \text{head}_\text{h})W^O, \quad (2)$$

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:

$$\text{MHA}(X) = \sum_{i=1}^{h} \text{head}_i W_i^O = \sum_{i=1}^{h} \text{Softmax}\left(\frac{X(W_i^Q {W_i^K}^\top)X^\top}{\sqrt{d_h}}\right)X(W_i^V W_i^O). \quad (3)$$

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.

3.2 Unilateral Weight Decomposition

Our UniDec 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:

$$\text{SVD}(W) = \tilde{W}(r) = US, \qquad L_I(W, r) = \|W - \tilde{W}(r)\|_F. \quad (4)$$

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 (b). 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:

$$\delta_i^{\text{QK}} = \mathbb{1}\left[\ell(W_i^Q, r_{QK}) \leq \ell(W_i^K, r_{QK})\right], \qquad \beta_i^{\text{VO}} = \mathbb{1}\left[\ell(W_i^V, r_{VO}) \leq \ell(W_i^O, r_{VO})\right]. \quad (5)$$

$$A_i^Q = \delta_i^{\text{QK}} U_i^Q + (1 - \delta_i^{\text{QK}}) W_i^Q {S_i^K}^\top, \qquad B_i^K = \delta_i^{\text{QK}} S_i^Q {W_i^K}^\top + (1 - \delta_i^{\text{QK}}) {U_i^K}^\top. \quad (6)$$

$$A_i^V = \beta_i^{\text{VO}} U_i^V + (1 - \beta_i^{\text{VO}}) W_i^V {S_i^O}^\top, B_i^O \quad = \beta_i^{\text{VO}} S_i^V {W_i^O}^\top + (1 - \beta_i^{\text{VO}}) {U_i^O}^\top, \quad (7)$$

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 UniDec is formulated as:

$$\text{MHA}(X) \approx \sum_{i=1}^{h} \text{Softmax}\left(\frac{X(A_i^Q B_i^K)X^\top}{\sqrt{r_{QK}}}\right)X(A_i^V B_i^O). \quad (8)$$

This formulation ensures that the model retains information on the intact side, yielding a more effective approximation than conventional per-weight decomposition. Compared to the naive per-weight SVD, our UniDec leads to a better performance-efficiency trade-off with respect to the number of parameters and FLOPs due to a more effective decomposition strategy.

Table 1: Comparison of the existing weight decomposition method and our UniDec 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
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
Unilateral Weight Decomposition (Ours)	50.3	**77.7**	81.1	31.4	**81.2**	57.3

Table 2: Performance comparison of the proposed UniDec on vision-language tasks. LLaVA1.5 is evaluated on widely used VQA (*i.e.*, VQAv2 and GQA) and multimodal evaluation (*i.e.*, MME) benchmarks. The reported parameters refer to the vision tower.

Method	LLaVA1.5-7B (Liu et al., 2024)				
	Params (M)	GFLOPs ↓	VQAv2 ↑	GQA ↑	MME ↑
Baseline	303.5	191.1	78.5	61.9	1504.6
Per-Weight Decomposition	272.1	173.0	69.6	56.9	1269.7
UniDec (Ours)	272.0	167.8	**74.7**	**59.8**	**1408.0**

4 Experiments

4.1 Experimental Settings

We evaluate on image classification (Deng et al., 2009) and vision-language tasks (Goyal et al., 2017; Hudson & Manning, 2019; Yin et al., 2024). We adopt diverse vision models (Dosovitskiy et al., 2020; Touvron et al., 2021; Liu et al., 2021; Touvron et al., 2022; Oquab et al., 2023) as baselines. Model compression was conducted on a single RTX 3090 GPU and the latency was measured with 224×224 resolution. Our approach does not require any fine-tuning process and compresses models by adjusting the decomposition ranks. The decomposition ranks were configured in a layer-wise hierarchical manner. In experiments, all decomposition methods start from the same checkpoints for fair comparisons.

4.2 UniDec for Large Vision Models

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

4.3 UniDec for Large Multi-modal Model

To further validate the effectiveness of the proposed UniDec, we conduct extensive experiments on vision-language tasks in Table 2. For LLaVA-1.5-7B, since the proposed method targets the vision attention modules, we apply our method to the vision tower. As shown in Table 2, UniDec achieves superior performance compared to the per-weight decomposition method under similar parameter and GFLOPs budgets. In particular, despite comparable reductions in parameters and computational cost, UniDec improves VQAv2 performance by 5.1% and GQA performance by 2.9% over the per-weight decomposition method. The results show that the proposed UniDec preserves task-relevant representations and generalizes well to diverse vision–language models.

Table 3: Performance comparison on downstream tasks. DINO (Zhang et al., 2022) and SETR (Zheng et al., 2021) are adopted as baselines for object detection and semantic segmentation, respectively.

Method	Object Detection			Semantic Segmentation		
	Params (M)	GFLOPs ↓	mAP ↑	Params (M)	GFLOPs ↓	mIoU ↑
Baseline	195.2	756.9	57.2	305.5	959.6	77.4
Per-Weight Decomposition	184.6	712.4	50.1	280.0	900.9	75.9
UniDec (Ours)	184.6	680.4	**55.4**	280.3	836.3	**76.9**

Table 4: Performance comparison of the proposed UniDec against various weight decomposition baselines on ImageNet-1K without fine-tuning.

Method	Params (M)	GFLOPs	Top-1 ↑	Latency ↓
Base model: DeiT-Base	86.6	33.7	81.8	30.7
ASVD (Yuan et al., 2023)	44.1	17.0	71.8	26.1
ASVD + **UniDec (Ours)**	44.1	17.0	**72.4**	**23.9**
SVD-LLM (Wang et al., 2024)	44.1	17.0	70.3	26.7
SVD-LLM + **UniDec (Ours)**	44.1	17.0	**72.2**	**25.1**
SVD-LLM v2 (Wang et al., 2025b)	44.1	17.0	70.5	26.3
SVD-LLM v2 + **UniDec (Ours)**	44.1	17.0	**72.8**	**25.2**

Table 5: Comparison to the per-weight decomposition method 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
UniDec (Ours)	**78.4**	**74.4**	**65.4**	**79.9**	**76.5**	**69.3**	**81.7**	**80.4**	**72.1**

4.4 Generalization to Downstream Tasks

To evaluate the generalization ability of the proposed UniDec on downstream tasks, we conduct experiments on object detection and semantic segmentation using COCO and Cityscapes datasets, respectively, in Table 3. We apply our UniDec and the per-weight decomposition method to DINO with a Swin Transformer backbone for object detection, and to SETR with a ViT backbone for semantic segmentation. As shown in Table 3, UniDec consistently outperforms the per-weight decomposition method on both downstream tasks. In particular, UniDec achieves a 5.3% higher mAP on object detection while maintaining comparable computational efficiency. These results demonstrate that UniDec generalizes well beyond image classification and can be effectively applied to diverse vision tasks and architectures.

4.5 UniDec with Diverse Decomposition Methods

Since the proposed UniDec is specifically designed for the vision attention decomposition, the practical deployment requires it to be used in conjunction with the existing weight decomposition methods applied to the MLP layers. Accordingly, we apply our UniDec to several representative LLM-based decomposition baselines and evaluate their performance on ImageNet-1K without fine-tuning. In Table 4, the experimental results demonstrate that UniDec can be generically integrated with the existing decomposition methods originally developed for LLMs, yielding consistent performance improvements. Under identical parameter and GFLOPs budgets, UniDec improves Top-1 accuracy by 1.9% when combined with SVD-LLM and by 2.3% when combined with SVD-LLM v2. These results indicate that UniDec is complementary to prior weight decomposition strategies and can be effectively incorporated into existing compression pipelines for vision models.

4.6 Comparison on Different Reduction Ratios

Table 5 compares UniDec with per-weight decomposition at different parameter reduction ratios by varying the ranks in MHA layers. UniDec consistently achieves higher accuracy across all reduction

Table 6: Performance of our UniDec on DeiT-Large and ViT-Large including fine-tuning (FT) results.

Method	DeiT3-Large			ViT-Large		
	Params in MHA	Top-1 ↑	Latency ↓	Params in MHA	Top-1 ↑	Latency ↓
Baseline	100.7	86.8	99.1	100.7	84.3	100.6
UniDec (w/o FT)	50.3	83.0	81.9	50.3	77.7	81.1
UniDec (w/ FT)	50.3	86.1	81.9	50.3	84.8	81.1

settings. At 60% reduction, it outperforms per-weight decomposition by 34.0 and 46.3 percentage points on DeiT-Small and DeiT-Small distilled, respectively, where per-weight decomposition suffers severe degradation. On DeiT-Base, UniDec remains robust under aggressive compression and incurs an accuracy drop of only 0.1 percentage points at 20% reduction. These results suggest that decomposing only one weight in each pair (W_i^Q, W_i^K) and (W_i^V, W_i^O) effectively mitigates information loss, providing a robust compression strategy for attention-based vision models.

4.7 Practical Operating Regime via Fine-Tuning

Although UniDec requires no training, brief fine-tuning can further recover accuracy lost to low-rank approximation. As shown in Table 6, just 10 epochs restore most of the lost accuracy while preserving the computational savings. On ViT-Large, fine-tuned UniDec achieves 84.8% Top-1 accuracy, surpassing the uncompressed baseline of 84.3%. These results demonstrate its effectiveness for practical model deployment.

4.8 Compatibility with Token Merging Strategies

A key advantage of the proposed UniDec is its orthogonality to other model compression paradigms, allowing it to be seamlessly integrated with alternative compression methods. To empirically validate this ecosystem compatibility, we evaluate the joint effect of combining our UniDec with ToMe, a representative token merging strategy, as displayed in Table 7. We specifically target a stringent compression regime where ToMe alone satisfies a low GFLOPs budget but suffers a severe performance degradation, dropping from the baseline accuracy of 81.8% down to 66.7%. Crucially, at the exact same GFLOPs constraint, integrating UniDec not only reduces the parameter count further (from 86.6M to 74.2M) but also delivers a substantial performance boost of 7.2%. This significant recovery demonstrates that UniDec acts as a powerful complement to sequence-reduction techniques, mitigating the information loss typical of extreme token pruning and highlighting its high practical utility for collaborative model optimization.

Table 7: Comparison of standalone ToMe (Bolya et al., 2022) versus our combined approach under an equivalent GFLOPs budget.

Method	Params (M)	GFLOPs	Top-1 Acc. (%)
DeiT-Base (Baseline)	86.6	33.7	81.8
ToMe (Bolya et al., 2022)	86.6	13.8	66.7
ToMe (Bolya et al., 2022) + **UniDec (Ours)**	**74.2**	13.8	**73.9**

4.9 Ablation on Selection Criterion

To assess the impact of the error metric used in our dynamic selection strategy, we conduct an ablation study comparing three candidate metrics for evaluating rank sensitivity: activation-aware scaling, spectral norm, and Frobenius norm. Table 8 presents the ablation results across various vision architectures under different rank constraints.

The empirical results reveal nuanced behaviors depending on the target rank. In highly constrained settings (*e.g.*, Rank 16 for DeiT/ViT or Rank 8 for Swin), no single metric universally dominates; for instance, the Frobenius norm outperforms others on DeiT-Base by a significant margin, whereas the spectral norm yields the best accuracy on DeiT-Large and ViT-Large. However, as the rank budget increases to more moderate regimes (*e.g.*, Rank 48 for DeiT/ViT and Rank 24 for Swin), the performance gap between the metrics narrows, with the Frobenius norm consistently achieving competitive or higher performance. Given its robust and stable performance across varying rank regimes, we adopt the Frobenius norm as the default selection criterion for the proposed UniDec.

Table 8: Ablation results for different error metric methods on various vision models.

Metric	DeiT-Base (Base head dim: 64)			DeiT-Large (Base head dim: 64)		
	Rank 16	Rank 32	Rank 48	Rank 16	Rank 32	Rank 48
Activation-aware	36.9	77.1	81.4	52.1	**83.6**	**86.3**
Spectral norm	37.6	77.3	81.4	**59.9**	83.3	84.9
Frobenius norm	**43.2**	**77.5**	**81.5**	56.6	83.0	85.1
	ViT-Large (Base head dim: 64)			Swin-Large (Base head dim: 32)		
	Rank 16	Rank 32	Rank 48	Rank 8	Rank 16	Rank 24
Activation-aware	23.1	75.9	80.8	21.6	81.1	**85.1**
Spectral norm	**33.6**	75.8	80.9	**22.4**	**81.5**	84.8
Frobenius norm	30.0	**77.7**	**81.2**	**22.4**	81.2	84.9

Table 9: Ablation study on the fixed selection cases and the dynamic selection for choosing less rank-sensitive elements within Q-K and V-O pairs. Results are reported under different parameter reduction ratios.

Method	Selected Elements				DeiT-Small			DeiT-Base		
	Q	K	V	O	20%	40%	60%	20%	40%	60%
Per-Weight Decomposition					75.7	67.2	31.4	80.7	79.0	71.4
UniDec w/ Fixed Selection (**Ours**)	✓		✓		75.4	65.9	39.7	81.4	79.5	61.8
	✓			✓	78.2	69.0	59.0	81.6	79.4	70.3
		✓	✓		75.8	71.3	46.5	81.5	80.4	65.6
		✓		✓	78.3	74.2	65.3	81.7	80.3	71.7
UniDec w/ Dynamic Selection (**Ours**)					**78.4**	**74.4**	**65.4**	**81.7**	**80.4**	**72.1**

Table 10: The ratio of Q/K and V/O selection on various vision models after the dynamic selection.

Models	Q / K Selection Ratio		V / O Selection Ratio	
	Q	K	V	O
DeiT-Base (Touvron et al., 2021)	20%	80%	10%	90%
DINOv2 (Oquab et al., 2023)	15%	85%	11%	88%
ViT-Large (Dosovitskiy et al., 2020)	30%	70%	7%	93%
Swin-Large (Liu et al., 2021)	17%	83%	30%	70%
DeiT3-Large (Touvron et al., 2022)	35%	65%	34%	66%

4.10 Effectiveness of Dynamic Selection

The proposed UniDec dynamically selects weights within each W^Q-W^K and W^V-W^O pair based on their sensitivity to low-rank approximation across individual heads. Rather than enforcing a uniform decomposition rule, this strategy adaptively determines which side of a pair should be factorized, thereby ensuring that the more stable component with respect to approximation error is preserved.

To verify its effectiveness, we conduct experiments by fixing decomposition cases for all four possible configurations of Q-K and V-O selections, and compared these cases with the dynamic selection strategy in Table 9. The results demonstrate that our dynamic approach more effectively mitigates information loss during weight decomposition. Specifically, compared to the least effective fixed selection case (*i.e.*, decomposing W^Q and W^V), the dynamic strategy achieves performance improvements of 10.3% at a 60% parameter reduction on DeiT-Base.

As the reduction rate increases, the fixed strategy of decomposing W^K and W^O yields performance close to the dynamic selection. This indicates that the low-rank approximation errors of W^K and W^O become smaller under higher reduction regimes, which aligns with the higher selection ratios for K and O observed in Table 10. However, as shown in Table 10, while K and O selections are more commonly observed, they do not exclusively dominate; Q and V still constitute a non-trivial proportion (up to 35% in DeiT3-Large) that varies noticeably across architectures. This variance underscores that the fixed selection strategy fails to capture the distinct structural sensitivities of different models, validating the necessity of our instance-adaptive dynamic selection.

Figure 2: Head-wise dynamic selection across layers. Results are shown for a reduced rank of 16 on DeiT-Base.

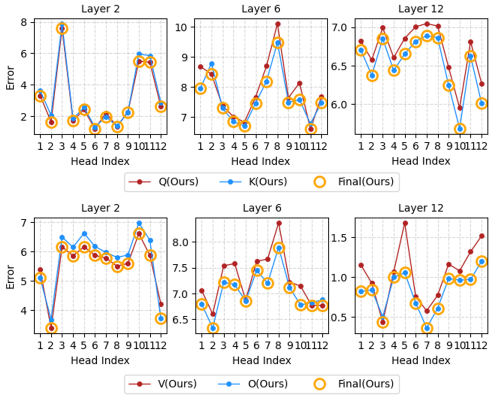


Figure 3: Selection maps per-layer/head on different vision models. The selection results vary across models and ranks.



4.11 Analysis of Head-wise Dynamic Selection

We employ a head-wise dynamic selection strategy for the W^Q-W^K and W^V-W^O pairs, where the side with the lower low-rank approximation error is chosen for decomposition. This design allows the decomposition process to adapt at the head level, ensuring that the most stable component in terms of approximation error is selected. To analyze the detailed selection behavior, we visualize the head-wise errors across layers in Figure 2. The results show that in the W^Q-W^K pair, W^K is more often selected, while in the W^V-W^O pair, W^O is more frequently chosen. Moreover, we visualize selection maps for different ranks and models in Figure 3. The resulting selections vary across models and ranks, but a consistent trend is observed that the Q and V tend to be selected more frequently in lower layers, and the K and O are selected more often overall. These results indicate that W^K and W^O exhibit lower sensitivity to low-rank approximation.

5 Conclusion

We proposed a novel singular value decomposition method for vision model compression, *UniDec*, 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. UniDec 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, especially under aggressive low-rank approximation. Experimental results demonstrated that our method effectively minimizes information loss compared to per-weight decomposition methods. Furthermore, UniDec can be universally integrated into the attention decomposition part of various decomposition methods, consistently achieving significant performance improvements and latency reductions.

6 Limitation

The proposed UniDec is designed upon a framework that pre-computes attention-related weights prior to inference to maximize efficiency. However, modern large language models primarily utilize Rotary Position Embedding (RoPE) within their Query (Q) and Key (K) projections to encode relative positional information. While RoPE can be mathematically reformulated into a linear matrix multiplication—making our decomposition method mechanistically applicable in theory—the resulting pre-computed transformation matrices become intrinsically bound to the specific sequence length of the input. Consequently, while UniDec can be seamlessly integrated into RoPE-based attention mechanisms under fixed sequence length conditions, applying it to scenarios with dynamic or variable sequence lengths remains challenging, as changing input lengths would invalidate the pre-computed weight structures.

AI USE STATEMENT

We used generative AI tools to support manuscript writing, LaTeX editing, and document formatting. We take full responsibility for the final content of this paper.

REFERENCES

Saleh Ashkboos, Maximilian L Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. Slicegpt: Compress large language models by deleting rows and columns. *arXiv preprint arXiv:2401.15024*, 2024.

Seyedarmin Azizi, Mahdi Nazemi, and Massoud Pedram. Memory-efficient vision transformers: An activation-aware mixed-rank compression strategy. In *European Conference on Computer Vision*, pp. 55–66. Springer, 2024.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Chi-Chih Chang, Wei-Cheng Lin, Chien-Yu Lin, Chong-Yan Chen, Yu-Fang Hu, Pei-Shuo Wang, Ning-Chi Huang, Luis Ceze, Mohamed S Abdelfattah, and Kai-Chiang Wu. Palu: Compressing kv-cache with low-rank projection. *arXiv preprint arXiv:2407.21118*, 2024a.

Chi-Chih Chang, Yuan-Yao Sung, Shixing Yu, Ning-Chi Huang, Diana Marculescu, and Kai-Chiang Wu. Flora: Fine-grained low-rank architecture search for vision transformer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2482–2491, 2024b.

Chi-Chih Chang, Chien-Yu Lin, Yash Akhauri, Wei-Cheng Lin, Kai-Chiang Wu, Luis Ceze, and Mohamed S Abdelfattah. xkv: Cross-layer svd for kv-cache compression. *arXiv preprint arXiv:2503.18893*, 2025.

Shih-Yang Liu1 Huck Yang Chein, Yi Wang Nai Chit Fung, Hongxu Yin Charbel Sakr, Saurav Muralidharan, Kwang-Ting Cheng1 Jan Kautz Yu, Chiang Frank Wang, and Pavlo Molchanov Min-Hung Chen. Eora: Training-free compensation for compressed llm with eigenspace low-rank approximation. *arXiv preprint arXiv:2410.21271*, 2024.

Patrick Chen, Hsiang-Fu Yu, Inderjit Dhillon, and Cho-Jui Hsieh. Drone: Data-aware low-rank compression for large nlp models. *Advances in neural information processing systems*, 34:29321–29334, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Compressing visual-linguistic model via knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1428–1438, 2021.

Shangqian Gao, Chi-Heng Lin, Ting Hua, Zheng Tang, Yilin Shen, Hongxia Jin, and Yen-Chang Hsu. Disp-llm: Dimension-independent structural pruning for large language models. *Advances in Neural Information Processing Systems*, 37:72219–72244, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Yangyang Guo, Guangzhi Wang, and Mohan Kankanhalli. Pela: Learning parameter-efficient models with low-rank approximation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15699–15709, 2024.

Yen-Chang Hsu, Ting Hua, Sungen Chang, Qian Lou, Yilin Shen, and Hongxia Jin. Language model compression with weighted low-rank factorization. *arXiv preprint arXiv:2207.00112*, 2022.

Xinhao Huang, You-Liang Huang, and Zeyi Wen. Sola: Leveraging soft activation sparsity and low-rank decomposition for large language model compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17494–17502, 2025.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Yixin Ji, Yang Xiang, Juntao Li, Wei Chen, Zhongyi Liu, Kehai Chen, and Min Zhang. Feature-based low-rank compression of large language models via bayesian optimization. *CoRR*, 2024.

Shibo Jie and Zhi-Hong Deng. Fact: Factor-tuning for lightweight adaptation on vision transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 1060–1068, 2023.

Jungi Lee, Wonbeom Lee, and Jaewoong Sim. Tender: Accelerating large language models via tensor decomposition and runtime requantization. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)*, pp. 1048–1062. IEEE, 2024.

Sunzhu Li, Peng Zhang, Guobing Gan, Xiuqing Lv, Benyou Wang, Junqiu Wei, and Xin Jiang. Hypoformer: Hybrid decomposition transformer for edge-friendly neural machine translation. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 7056–7068, 2022.

Yixiao Li, Yifan Yu, Qingru Zhang, Chen Liang, Pengcheng He, Weizhu Chen, and Tuo Zhao. Losparse: Structured compression of large language models based on low-rank and sparse approximation. In *International Conference on Machine Learning*, pp. 20336–20350. PMLR, 2023.

Chi-Heng Lin, Shangqian Gao, James Seale Smith, Abhishek Patel, Shikhar Tuli, Yilin Shen, Hongxia Jin, and Yen-Chang Hsu. Modegpt: Modular decomposition for large language model compression. *arXiv preprint arXiv:2408.09632*, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Chengtao Lv, Hong Chen, Jinyang Guo, Yifu Ding, and Xianglong Liu. Ptq4sam: Post-training quantization for segment anything. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 15941–15951, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Sergey Pletenev, Viktoriia Chekalina, Daniil Moskovskiy, Mikhail Seleznev, Sergey Zagoruyko, and Alexander Panchenko. 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.

Bosheng Qin, Juncheng Li, Siliang Tang, and Yueting Zhuang. Dba: Efficient transformer with dynamic bilinear low-rank attention. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

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

Yuzhang Shang, Zhihang Yuan, Qiang Wu, and Zhen Dong. Pb-llm: Partially binarized large language models. *arXiv preprint arXiv:2310.00034*, 2023.

Moritz Thoma, Jorge Villasante, Emad Aghajanzadeh, Shambhavi Balamuthu Sampath, Pierpaolo Mori, Maximilian Groetzinger, Daniil Dylkin, Manoj-Rohit Vemparala, Nael Fasfous, Alexander Frickenstein, 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.

Jiayi Tian, Ryan Solgi, Jinming Lu, Yifan Yang, Hai Li, and Zheng Zhang. Flat-llm: Fine-grained low-rank activation space transformation for large language model compression. *arXiv preprint arXiv:2505.23966*, 2025.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

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

Lokesh Veeramacheneni, Moritz Wolter, Hilde Kuehne, and Juergen Gall. Canonical rank adaptation: An efficient fine-tuning strategy for vision transformers. In *Forty-second International Conference on Machine Learning*, 2025.

Qinsi Wang, Jinghan Ke, Masayoshi Tomizuka, Yiran Chen, Kurt Keutzer, and Chenfeng Xu. Dobi-svd: Differentiable svd for llm compression and some new perspectives. *arXiv preprint arXiv:2502.02723*, 2025a.

Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 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.

Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. Svd-llm: Truncation-aware singular value decomposition for large language model compression. *arXiv preprint arXiv:2403.07378*, 2024.

Xin Wang, Samiul Alam, Zhongwei Wan, Hui Shen, and Mi Zhang. Svd-llm v2: Optimizing singular value truncation for large language model compression. *arXiv preprint arXiv:2503.12340*, 2025b.

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

Jinqi Xiao, Miao Yin, Yu Gong, Xiao Zang, Jian Ren, and Bo Yuan. Comcat: towards efficient compression and customization of attention-based vision models. *arXiv preprint arXiv:2305.17235*, 2023.

Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Yuanlin Duan, Wenqi Jia, Miao Yin, Yu Cheng, and Bo Yuan. Moe-i2: Compressing mixture of experts models through inter-expert pruning and intra-expert low-rank decomposition. *arXiv preprint arXiv:2411.01016*, 2024a.

Huanrui Yang, Hongxu Yin, Maying Shen, Pavlo Molchanov, Hai Li, and Jan Kautz. Global vision transformer pruning with hessian-aware saliency. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18547–18557, 2023.

Yifan Yang, Jiajun Zhou, Ngai Wong, and Zheng Zhang. Loretta: Low-rank economic tensor-train adaptation for ultra-low-parameter fine-tuning of large language models. *arXiv preprint arXiv:2402.11417*, 2024b.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

Hao Yu and Jianxin Wu. 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.

Zhihang Yuan, Yuzhang Shang, Yue Song, Qiang Wu, Yan Yan, and Guangyu Sun. Asvd: Activation-aware singular value decomposition for compressing large language models. *arXiv preprint arXiv:2312.05821*, 2023.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

Borui Zhao, Renjie Song, and Jiajun Liang. Cumulative spatial knowledge distillation for vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6146–6155, 2023.

Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6881–6890, 2021.

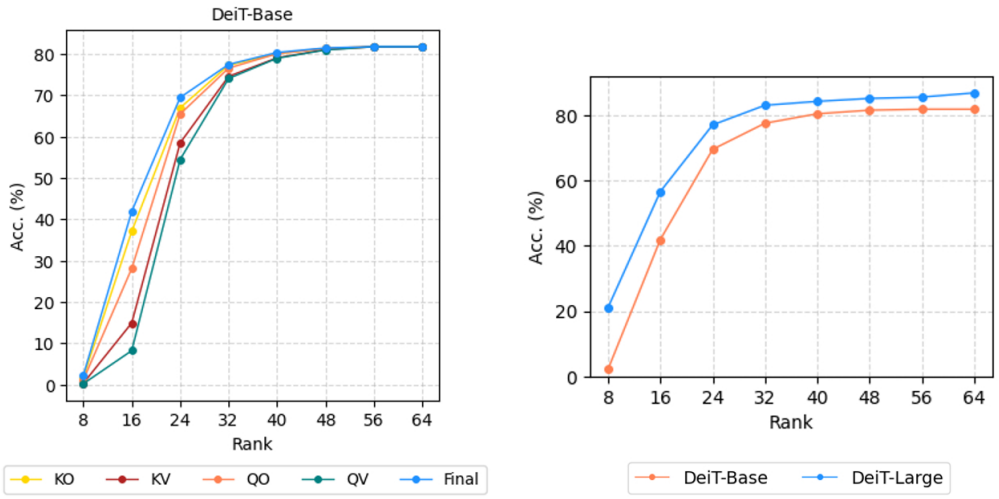


Figure 4: (Left) Comparison of rank sensitivity. Dynamic selection shows better robustness than the fixed cases. (Right) Rank–performance curves for DeiT-Base and DeiT-Large.

Table 11: Wall-clock time comparison for dynamic selection. Each time is the cumulative time for all layers. Results are measured on a single RTX 3090 GPU.

Method	DINOv2	ViT-Large	DeiT-Large	Swin-Large
w/o dynamic selection (sec)	0.6	1.3	1.1	1.3
w/ dynamic selection (sec)	1.2	2.6	2.4	2.6

A Appendix

A.1 Analysis of Rank Sensitivity

In Figure 4, we conduct a comprehensive rank sensitivity analysis by plotting rank–performance trade-offs for fixed-case ablations alongside different model scales. The fixed-case evaluation on the left highlights that our dynamic selection mechanism consistently surpasses all static rank choices across all rank settings. Notably, this advantage becomes increasingly pronounced in lower-rank regimes, underscoring the necessity of dynamic rank allocation when aggressive compression is applied. When examining model scalability on the right, both DeiT-Base and DeiT-Large display remarkably aligned behavior: while higher ranks yield diminishing performance gains, reducing the rank below 24–32 leads to a steep decline in accuracy. These results suggest that the 24–32 rank boundary serves as an intrinsic capacity floor for effective vision transformer representation.

A.2 Wall-clock Time Analysis

Since the dynamic selection process can be precomputed, this selection strategy does not incur additional computational overhead at inference time compared to the fixed selection strategy. Despite this fact, we measure the cumulative wall-clock time across all layers in Table 11 to quantify the preprocessing cost required for the adaptive choice. The results show that the dynamic selection method incurs only a minor time increment compared to the fixed selection baseline. Specifically, the absolute overhead for evaluating rank sensitivity ranges from 0.6 seconds (DINOv2) to 1.3 seconds (ViT-Large, DeiT-Large, and Swin-Large). This overhead is a one-time cost incurred during the offline compression phase and does not affect the inference efficiency of UniDec.