Prune Once: Retraining-Free Task-Agnostic Pruning for Vision-Language Models

Minseok Kang[1*], Hyunwoo Kim[1*], Chanyoung Kim[2], Minwoo Kim[1], Jaekoo Lee[3†], and Dahuin Jung[1†]

[1] Department of Artificial Intelligence, Chung-Ang University, Seoul, Korea
[2] School of Computer Science and Engineering, Soongsil University, Seoul, Korea
[3] Department of Computer Science, Kookmin University, Seoul, Korea
dahuinjung@cau.ac.kr

Abstract. Vision-language models (VLMs) have achieved remarkable generalization across diverse multimodal tasks through large-scale pretraining, yet their rapidly increasing computational and memory requirements pose significant challenges for deployment in constrained environments. Existing pruning strategies often depend on task-specific criteria or LLM-oriented importance measures, making them unsuitable for task-agnostic pruning, where no task-specific samples are available at pruning time and the pruned model remains broadly applicable. We introduce a retraining-free VLM pruning framework called PORTA that derives a task- and modality-agnostic importance formulation based on activation variation, estimated from generic calibration data, which reliably captures feature-level representation utility across modalities. PORTA further incorporates an adaptive sparsity allocation mechanism that assigns layer-wise pruning ratios based on output feature variability, avoiding the limitations of uniform sparsity and reducing performance degradation at high compression levels. Extensive experiments across VLM architectures, such as CLIP, BLIP, and Qwen2-VL, demonstrate that PORTA achieves competitive downstream performance under high sparsity without requiring any retraining, supporting efficient VLM compression. Code is available at https://github.com/cau-hai-lab/PORTA.git.

1 Introduction

Vision-language models (VLMs) [21, 22, 28] have established a unified framework for joint visual and textual representation learning, enabling a broad spectrum of multimodal understanding tasks. Through large-scale pretraining on web-scale image–text pairs, these models effectively capture cross-modal correspondences and exhibit strong generalization across downstream tasks such as retrieval [6] and VQA [2]. Pioneering frameworks such as CLIP [28] and BLIP [22] exemplify this progress, leveraging web-scale data to acquire transferable multimodal

*Equal Contribution , †Corresponding Authors

features that enable zero-shot adaptation to new domains and tasks, thereby unifying multiple vision-language capabilities within a single model.

Despite their impressive performance, the ever-increasing scale of VLMs in terms of parameters and computation has rendered their deployment highly challenging in resource-constrained environments such as edge devices and real-time systems. To mitigate these limitations, a broad spectrum of model compression techniques—such as quantization [12, 39], knowledge distillation [37], and pruning—have been actively investigated. Among these, pruning [1, 9, 16, 20, 29, 30, 32, 33, 38] has emerged as a prominent method that reduces network complexity by removing parameters with low contribution to task performance, leading to reduced computational cost while maintaining accuracy.

Task-agnostic pruning for VLMs —where a model is pruned once and reused across diverse downstream tasks without re-pruning or retraining, as illustrated in Fig. 1—remains unexplored despite its practical importance. Most existing pruning methods rely on task-specific criteria, requiring re-pruning or retraining whenever the downstream task changes, which introduces non-negligible overhead [5]. In particular, in VLM pruning, many approaches adopt LLM-based weight-importance measures such as Wanda [30] and SparseGPT [10]. However, recent work shows that the resulting importance estimates can be sensitive to the choice of calibration data [39]. Consequently, in task-agnostic settings—where task-specific calibration samples are difficult to obtain—importance estimation may become unstable, limiting the applicability of such methods. Moreover, Wanda is motivated by activation outlier patterns observed in LLMs [30]. As illustrated in Fig. 2 (a), these patterns do not manifest in the same manner in vision models, which may reduce their effectiveness when applied to VLMs. This observation suggests that pruning criteria derived from activation characteristics in LLMs may not generalize uniformly across the vision and language components of VLMs. In particular, when importance estimation relies heavily on modality-specific activation distribution properties, it may introduce modality bias, causing pruning to be disproportionately concentrated in either the vision or language branch. This issue becomes more pronounced in task-agnostic scenarios, where the absence of task-specific calibration signals prevents effective mitigation of such bias.

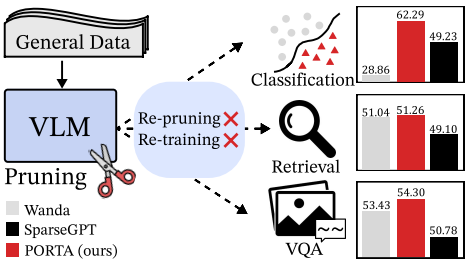

Fig. 1: Illustration of PORTA's task- and modality-agnostic pruning paradigm: using only general data, a VLM is pruned once—without any re-pruning or retraining—and directly applied to diverse downstream tasks. Comparison across three tasks—classification (accuracy), retrieval (TR@5), and VQA (Avg)—under high sparsity. PORTA delivers uniformly strong results across all settings, outperforming Wanda and SparseGPT.

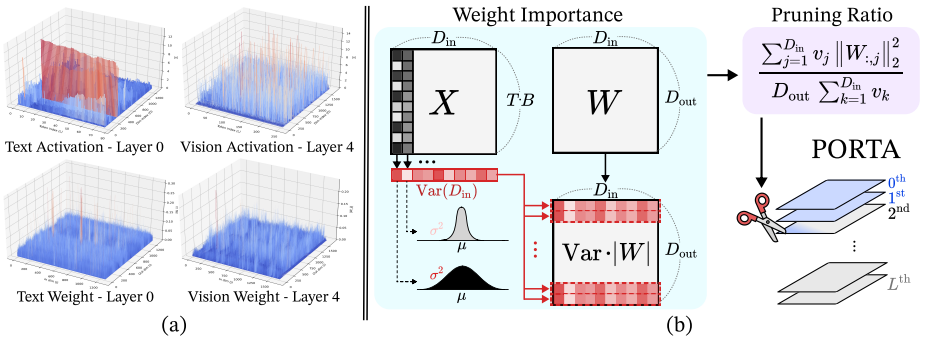

Fig. 2: Modality-specific activation characteristics and PORTA overview. The left component, (a) Text and vision branches exhibit clearly different activation patterns across layers. (b) PORTA overview: Var·$|W|$ for weight importance and a pruning-ratio score for layer-wise sparsity allocation.

To address this bias, we propose PORTA (Prune-Once: Retraining-free Task-Agnostic pruning), which does not simply replace existing weight-importance statistics but introduces a new importance formulation that can reflect the characteristics of VLMs. PORTA utilizes activation variation, which remains comparatively stable across modalities, to capture each feature's representation utility and construct a task- and modality-agnostic importance formulation. Specifically, PORTA derives feature importance from the variation of activations along the feature dimension, interpreting broader fluctuations as evidence of stronger engagement across diverse inputs and higher representation utility. This formulation mitigates modality-induced bias, maintains coherent feature-space behavior by grounding pruning decisions in the model's representational behavior, and enables seamless transfer across heterogeneous tasks without re-pruning, as its importance estimates are robust to the choice of calibration data, achieving improved efficiency across modalities. Concretely, PORTA incorporates two components to enable layer-wise sparsity allocation: a scoring criterion for selecting important features and a ratio mechanism that adaptively allocates layer-wise sparsity across network blocks. The scoring criterion estimates feature importance from the activation variation along the feature axis—low-variability features that react only to specific contexts are pruned, while high-variability features with broader responsiveness are retained. The ratio mechanism assigns different sparsity levels to each layer according to the variability of its output feature representation, measured from activation variation, thereby alleviating the limitation of layer-uniform pruning. Layers exhibiting higher representation utilization are assigned lower sparsity, thereby mitigating performance cliffs through selective pruning of less contributive features.

As shown in Fig. 1, across extensive zero-shot evaluations, PORTA consistently delivers strong pruning performance on a wide range of VLM architectures and tasks. On CLIP, PORTA achieves substantial gains over Wanda, SparseGPT, ECoFLaP [31], and Multiflow [9] in image classification and image–text retrieval, particularly at high sparsity levels where alternative methods

exhibit severe degradation. Moreover, PORTA generalizes beyond CLIP-based contrastive VLMs to generative multimodal architectures such as Qwen2-VL [36], while preserving strong downstream performance on VQA [2] even after pruning. Our experiments confirm that PORTA maintains strong downstream performance across architectures, tasks, and sparsity levels, demonstrating that task- and modality-agnostic pruning can be achieved without re-pruning, as the proposed importance estimation remains robust to the choice of calibration data. The main contributions of this work are summarized as follows:

- We propose PORTA, a prune-once, retraining-free pruning framework for VLMs. PORTA prunes a pretrained model once and transfers it across diverse downstream tasks without any additional training.
- We introduce a modality-agnostic importance criterion and an adaptive sparsity allocation mechanism. The criterion quantifies feature salience via activation-based statistical variability, reflecting each feature's contribution to representational capacity, while sparsity ratios are adaptively assigned based on the variability of the output feature space at each layer.
- We demonstrate that PORTA achieves robust performance across multiple VLM architectures and sparsity levels. In particular, PORTA maintains competitive accuracy compared with existing pruning methods even under high sparsity, validating its robustness and efficiency in large-scale multimodal compression.

2 Related Works

2.1 Vision-Language Models

Vision-language models (VLMs) that jointly learn visual and textual representations have been actively explored in recent years [21, 22, 24, 28]. Among them, CLIP [28] demonstrates that a single model can be broadly applied to various vision-language tasks such as classification, retrieval, and VQA, by aligning visual and textual representations in a shared embedding space through contrastive learning on large-scale image–text pairs. Building on this paradigm, BLIP [22] connects a pretrained vision encoder with a language model to enable more general-purpose vision-language generation and understanding. Recent studies further expand model capacity to capture more complex multimodal information, yet the resulting increase in parameters and computational demand poses challenges for deployment in resource-constrained environments.

2.2 Post-Training Pruning

As models continue to scale, the computational cost of iterative pruning has become prohibitive, motivating the adoption of one-shot pruning as a more practical alternative. Representative approaches include SparseGPT [10] and Wanda [30], which leverages the emergent large-magnitude activations observed

in LLMs. However, when applied to VLMs, these strategies reveal notable limitations. Recent findings [39] indicate that both SparseGPT and Wanda are highly sensitive to the selection of calibration data. In task-agnostic VLM scenarios—where task-specific calibration data cannot be utilized—this sensitivity results in suboptimal pruning outcomes. Furthermore, Wanda's reliance on activation norms makes it sensitive to the differing activation magnitudes of VLMs, making it less appropriate for multimodal pruning.

Recent studies have attempted to design pruning strategies specifically for VLMs, such as Multiflow [9] and ECoFLaP [31]. Multiflow focuses on information flow by leveraging modality-specific weight distributions for task-agnostic pruning, whereas ECoFLaP introduces an efficient layer-wise ratio estimation method using zeroth-order gradients to reduce computational overhead. Nonetheless, Multiflow still requires task-specific fine-tuning, incurring additional cost, and ECoFLaP does not introduce its own importance metric; instead, it adopts Wanda, whose assumptions are rooted in LLM behavior and are not directly aligned with VLMs.

2.3 Task-Agnostic Pruning

In continual learning [14] and few-shot learning [15], task-agnostic approaches aim to acquire representations or update rules that generalize to unseen tasks, emphasizing robustness to task shifts rather than reliance on task-specific supervision. Beyond learning, the notion of "task-agnostic" has gained renewed importance in the context of model compression for large models [35, 40, 42]. Among them, Multiflow [9] highlights that VLMs are typically deployed across heterogeneous downstream tasks, and that using task-specific calibration data for pruning undermines generalization. Accordingly, Multiflow formalizes task-agnostic pruning as performing pruning without access to downstream labels or task-specific samples, ensuring that the pruned model remains broadly usable across retrieval, classification, and generation tasks. In this landscape, PORTA advances task-agnostic pruning by performing pruning without any form of retraining, achieving stable performance across modalities and tasks.

3 PORTA

In this section, we present the motivation and key components of PORTA, our pruning framework designed for vision–language models (VLMs). PORTA aims to provide a pruning mechanism that remains robust to variations in calibration data and modality-specific activation characteristics—two limitations that hinder the reliability of existing VLM pruning approaches. To address these limitations, PORTA introduces two components. First, we develop a novel importance formulation that mitigates input-distribution sensitivity in multimodal pruning settings by leveraging feature activation variance, enabling a reliable assessment of each feature's representation utility. Second, we propose a layer-wise pruning ratio estimation strategy that accounts for architectural heterogeneity in

Table 1: Layer-wise max–min dispersion of activation statistics under magnitude- and variance-based measures. Variance exhibits reduced modality sensitivity.

Method	Module	L1	L20	L30
Mag.	Text	3.44	7.11	7.07
Mag.	Vision	4.38	4.85	3.73
Var.	Text	3.67	4.48	4.61
Var.	Vision	3.73	4.47	4.11

Table 2: Zero-shot retrieval accuracy after removing a percentage of features based on variance ranking.

Removed Features	5%	10%	15%
High-Variance	0.02	0.02	0.02
Random-Variance	67.00	63.62	47.60
Low-Variance	68.34	67.88	67.28

VLMs by evaluating the variability of layer output feature spaces. An overview of PORTA is provided in Fig. 2 (b).

3.1 Variance as a Modality-Agnostic Importance Score

Existing pruning methods predominantly rely on activation magnitude as the primary criterion for importance estimation. This assumption is largely motivated by activation outlier patterns observed in large language models, where dimensions exhibiting extreme responses are often considered highly influential. However, such magnitude-based assumptions may not directly transfer to multimodal architectures such as VLMs. In VLMs, the activation scale and distribution characteristics differ substantially between the vision and language encoders. In particular, pronounced outlier patterns commonly observed in text representations are less evident in vision features. Such modality-dependent differences can also be observed in Fig. 2 (a). As a result, magnitude-based importance estimation may introduce modality-dependent bias, potentially leading to imbalanced pruning across modalities. To analyze modality sensitivity, we examine magnitude- and variance-based activation statistics, and then analyze the relationship between feature variance and representation utility.

Modality Sensitivity of Activation Statistics. We analyze the modality sensitivity of activation statistics by measuring their dispersion across layers and modalities. For each layer, we compute two quantities: (i) the range of dimension-wise mean activations, defined as the difference between the maximum and minimum mean activation values across feature dimensions; and (ii) the range of dimension-wise activation variances, defined as the difference between the maximum and minimum variance values across feature dimensions.

Tab. 1 reports these max–min ranges under magnitude- and variance-based statistics. As shown, magnitude-based statistics exhibit substantial variation across layers and between modalities, reflecting strong sensitivity to modality-specific activation scales. In contrast, variance-based statistics demonstrate comparatively stable dispersion across both vision and language encoders, indicating reduced modality dependence. These findings suggest that variance provides a more stable and modality-agnostic signal than magnitude in multimodal settings, motivating its use as the foundation of our importance formulation.

Representation Utility and Feature Variance. We consider feature variance as an alternative indicator of representation utility. Intuitively, features with larger variance tend to respond broadly across tokens or patches and thus capture generalizable structures, whereas low-variance features often encode localized or input-specific patterns.

To empirically validate the relationship between feature variance and representation utility, we conduct an axis-removal analysis based on feature variance in Tab. 2. Specifically, we compare three removal schemes: (i) High-Variance features, (ii) Random features, and (iii) Low-Variance features.

For each scheme, we remove a certain percentage of feature dimensions according to the variance ranking while keeping all other model parameters fixed, and evaluate zero-shot retrieval performance.

Across all tested removal ratios, removing high-variance features causes catastrophic performance degradation, whereas removing low-variance features results in minimal accuracy drop. These results indicate that high-variance features carry most of the representational utility, validating variance as a meaningful importance signal.

The results indicate that low-variance features can be removed with minimal impact on performance and variance provides a robust and modality-agnostic basis for importance estimation in multimodal pruning. Unlike magnitude-based criteria that focus on extreme responses, variance reflects the breadth of representation utilization across inputs, providing a balanced measure of contribution.

Based on the above analysis, feature variance is incorporated into the weight tensor following the standard weight-pruning paradigm [13, 30].

3.2 Weight Importance

We quantify representation utility for each layer by measuring the variance of its input features along the token/patch dimension. Let the input activations to layer l be $X \in \mathbb{R}^{B\times T\times D_{\text{in}}}$ and the corresponding weight matrix be $W \in \mathbb{R}^{D_{\text{out}}\times D_{\text{in}}}$. For each input feature j, we compute its variability as

$$v_j = \text{Var}_{b,t}(X_{b,t,j}), \qquad j = 1, \ldots, D_{\text{in}}, \tag{1}$$

where v_j serves as a measure of representation utility: features exhibiting low variance respond only to a narrow subset of input tokens or patches and thus encode highly specific, localized patterns, whereas high-variance features participate broadly across inputs and therefore capture generalizable structure.

To incorporate representation utility into pruning, we combine feature variance with weight magnitude. The variance vector v is broadcast along the output dimension and multiplied with the absolute weight matrix, yielding the final importance score

$$S_{i,j} = v_j \cdot |W_{i,j}|, \qquad i = 1, \ldots, D_{\text{out}},\ j = 1, \ldots, D_{\text{in}}, \tag{2}$$

where $S_{i,j}$ reflects both the scale of each parameter and the extent to which the corresponding input feature contributes to the model's representational capacity.

3.3 Pruning Ratio

Different layers within deep models are known to play distinct roles and contribute unequally to the resulting representations [19, 34, 41]. In VLMs, this variation is further influenced by their structurally distinct architectural components and modality-specific fusion structures, which lead layers to participate differently in forming multimodal features. Since such architectures do not impose a fixed modality weighting or uniform layer structure, assigning a consistent pruning ratio across layers requires a measure that is both modality- and task-agnostic. To achieve a balanced allocation of pruning across diverse structures, PORTA measures each layer output's feature-space utilization using its output variance, which serves as an indicator of how much representational capacity the layer occupies.

To define and introduce the pruning-ratio score for each layer, we first establish the notation. Let $X \in \mathbb{R}^{T \times D_{\text{in}}}$ denote the input tokens, where each row x_t represents a token, and let $W \in \mathbb{R}^{D_{\text{out}} \times D_{\text{in}}}$ be the weight matrix of a linear layer. The layer output is given by

$$Y = XW^\top \in \mathbb{R}^{T \times D_{\text{out}}}. \tag{3}$$

Denoting the o-th row of W by $w_o \in \mathbb{R}^{1 \times D_{\text{in}}}$, the o-th output channel (the o-th column of Y) can be expressed as

$$y_{:,o} = X w_o^\top. \tag{4}$$

For analytical simplicity, we consider each input token x_t is drawn from a zero-mean distribution with input covariance $\Sigma_X = \mathbb{E}[x_t^\top x_t]$, the variance of the o-th output channel across the token dimension then becomes

$$\begin{aligned}
\text{Var}[y_{:,o}] &\approx \mathbb{E}_{x_t}\left[(x_t w_o^\top)^2\right] && (5)\\
&= \mathbb{E}_{x_t}\left[x_t w_o^\top w_o x_t^\top\right] && (6)\\
&= \mathbb{E}_{x_t}\left[\text{Tr}(w_o^\top w_o x_t^\top x_t)\right] && (7)\\
&= \text{Tr}\big(w_o^\top w_o \, \Sigma_X\big) \;=\; w_o \Sigma_X w_o^\top. && (8)
\end{aligned}$$

We propose a layer-wise score for determining pruning ratios in PORTA. To eliminate scale effects and ensure fair comparison across layers, we first normalize the input covariance so that only its directional structure is retained:

$$\widehat{\Sigma}_X \;\coloneqq\; \frac{\Sigma_X}{\text{Tr}(\Sigma_X)}. \tag{9}$$

Based on normalized covariance, the pruning ratio for each layer is determined using the following score:

$$s_{\text{layer}} \;\coloneqq\; \frac{1}{D_{\text{out}}} \sum_{o=1}^{D_{\text{out}}} w_o \, \widehat{\Sigma}_X \, w_o^\top, \tag{10}$$

Algorithm 1 Layer-wise Pruning Ratio Allocation

Require: $\{s_i\}_{i\in\mathcal{L}},\ S\in(0,1),\ \alpha>0,\ \{w_i\}_{i\in\mathcal{L}}$
Ensure: $\{r_i\}_{i\in\mathcal{L}}$
$T \leftarrow \sum_{i\in\mathcal{L}} s_i$ ▷ sum over layer importance scores
$p_i \leftarrow s_i/T$, for all $i\in\mathcal{L}$ ▷ importance proportion
$u_i \leftarrow (1-p_i)^{\alpha}$, for all $i\in\mathcal{L}$ ▷ pruning score with strength α
$W \leftarrow \sum_i w_i$
$\bar{u} \leftarrow (\sum_i w_i u_i)/W$ ▷ weight-adjusted mean pruning score
$C \leftarrow S/\bar{u}$
$r_i \leftarrow C\,u_i$, for all $i\in\mathcal{L}$ ▷ per-layer pruning ratio
return $\{r_i\}_{i\in\mathcal{L}}$

where averaging over output channels ensures consistency across layers with different output dimensionalities.

By stacking all output weight vectors w_o into a matrix $W=[w_1;\ldots;w_{D_{\text{out}}}]$, we equivalently write Eq. 10 as follows:

$$s_{\text{layer}} = \frac{1}{D_{\text{out}}}\,\text{Tr}\big(W^\top W\,\widehat{\Sigma}_X\big) = \frac{1}{D_{\text{out}}}\sum_{j=1}^{D_{\text{in}}}\big\|\big(W\,\widehat{\Sigma}_X^{1/2}\big)_{:,j}\big\|_2^2. \tag{11}$$

As illustrated in Fig. 2 (b), the formulation can be interpreted as the average squared ℓ_2 norm of the input-direction feature vectors of W projected onto the input feature covariance space. The layer output feature variance score reflects the joint contribution of the input feature normalized covariance $\widehat{\Sigma}_X$and the ℓ_2 magnitude of the weight vectors along each input direction.

We approximate the covariance Σ_X with its diagonal for computational and memory efficiency, $\Sigma_X \approx \text{diag}(v_1,\ldots,v_{D_{\text{in}}})$ with $v_j=\text{Var}[x_{t,j}]$. In variance-only case, s_{layer} reduces to

$$s_{\text{layer}} \;=\; \frac{\sum_{j=1}^{D_{\text{in}}} v_j\,\big\|W_{:,j}\big\|_2^2}{D_{\text{out}}\,\sum_{k=1}^{D_{\text{in}}} v_k}, \qquad v_j=\text{Var}[x_{t,j}], \tag{12}$$

where each column vector $W_{:,j}$ quantifies the influence of the j-th input feature on the output channels, while the variance term v_j measures the representation utility of the input feature. The resulting score s_{layer} represents a normalized, variance-weighted average of $\|W_{:,j}\|_2^2$, assigning larger contribution to input dimensions that both exhibit higher utility and exert stronger effect on the layer's output. Layers whose output representations rely heavily on such high-utility directions receive larger scores and are therefore allocated lower pruning ratios. The computed s_{layer} values are further normalized across layers to satisfy a target global sparsity, as detailed in Algorithm 1.

4 Experiments

In this section, we conduct zero-shot evaluations of PORTA across three task categories—image classification, image–text retrieval, and VQA—using three repre-

Table 3: Zero-shot evaluation on image–text retrieval and image classification using CLIP pruned at 55%, 60%, and 65% sparsity. PORTA consistently achieves strong performance across both retrieval (MSCOCO) and classification benchmarks, particularly under high sparsity, where alternative pruning methods exhibit substantial degradation. C10: CIFAR-10, C100: CIFAR-100, IN: ImageNet, F102: Flowers102.

Sparsity	Method	Retrieval				Classification			
		MSCOCO				C10	C100	IN	F102
		TR@1	TR@5	IR@1	IR@5	Accuracy			
0%	**Dense**	68.56	87.70	52.28	76.09	97.04	87.50	78.45	80.19
55%	Wanda	**66.30**	86.44	48.62	73.05	**96.12**	79.92	69.50	67.72
	SparseGPT	65.78	86.44	48.57	**73.29**	95.07	79.95	68.61	65.16
	ECoFLaP	51.14	76.04	33.14	58.44	80.69	51.79	43.11	30.00
	Multiflow	0.02	0.12	0.02	0.12	9.25	1.23	0.16	1.50
	PORTA	66.28	**87.00**	**48.70**	73.02	95.58	**80.27**	**70.29**	**68.73**
60%	Wanda	59.78	82.82	42.55	68.01	93.20	70.28	57.35	51.48
	SparseGPT	57.52	81.88	41.77	66.97	**95.16**	74.23	55.26	41.37
	ECoFLaP	35.44	60.56	19.12	39.46	57.98	32.60	25.73	14.21
	Multiflow	0.04	0.16	0.01	0.08	9.79	1.10	0.08	0.49
	PORTA	**60.14**	**83.04**	**43.11**	**68.18**	95.06	**75.59**	**59.66**	**55.50**
65%	Wanda	27.30	51.04	16.96	35.61	69.81	28.86	30.06	17.30
	SparseGPT	25.10	49.10	15.93	35.00	86.35	49.23	25.96	12.09
	ECoFLaP	13.60	29.62	7.22	18.32	33.18	16.70	10.20	6.66
	Multiflow	0.02	0.12	0.02	0.10	8.90	0.96	0.12	0.81
	PORTA	**28.14**	**51.26**	**17.31**	**37.32**	**90.49**	**62.29**	**31.38**	**18.27**

sentative VLMs: OpenAI CLIP-ViT-bigG, BLIP-base [22], and Qwen2-VL [36]. All evaluations are performed in a fully zero-shot setting using publicly released pretrained checkpoints.

4.1 Experimental Setup

We evaluate PORTA under zero-shot settings across three task families: image classification, image–text retrieval, and VQA. For classification, we use four benchmarks—CIFAR-10 [18], CIFAR-100 [18], ImageNet-1K [7], and Oxford Flower-102 [26]—to assess performance across datasets with varying granularity and scale. Image–text retrieval experiments are conducted on MSCOCO [23] and Flickr30K [27]. For MSCOCO, we follow the Karpathy split and report Recall@K for both image-to-text and text-to-image retrieval on the Karpathy test split, while Flickr30K results are obtained using its official test split. For VQA, we use ScienceQA [25] and report test-split accuracy across its subject, context-modality, and grade categories, along with the overall average.

We assess PORTA on three representative VLMs using publicly available pretrained checkpoints without any fine-tuning: OpenAI CLIP-ViT-bigG and BLIP-base for classification and retrieval, and Qwen2-VL [36] for VQA. For comparison, we include four pruning baselines: Wanda [30], which combines activation statistics with weight magnitudes; SparseGPT [10], which uses Hessian-

Table 4: Zero-shot image–text retrieval results on Flickr30k using BLIP at 45% sparsity. PORTA attains the highest mean performance among pruning baselines.

Sparsity	Method	Image-Text Retrieval								
		Flickr30k								
		TR@1	TR@5	TR@10	TR@mean	IR@1	IR@5	IR@10	IR@mean	Mean
0%	**Dense**	83.10	96.60	97.90	92.53	78.48	93.82	96.52	89.61	91.07
45%	Wanda	82.20	95.70	98.10	92.00	74.40	92.82	95.66	87.63	89.81
	SparseGPT	81.70	95.10	**98.20**	91.67	**75.86**	93.02	**95.82**	**88.23**	89.95
	ECoFLaP	79.20	93.20	96.40	89.60	69.78	90.50	94.66	84.98	87.29
	Multiflow	79.50	94.20	97.00	90.43	74.48	92.50	95.48	87.49	88.96
	PORTA	**83.00**	**96.00**	98.00	**92.33**	75.62	**93.10**	**95.82**	88.18	**90.26**

Table 5: Zero-shot VQA results on ScienceQA at 50% sparsity with Qwen2-VL. PORTA achieves the highest average accuracy among pruning baselines.

Sparsity	Method	Subject			Context Modality			Grade		Average
		NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
0%	**Dense**	60.66	67.27	60.00	62.37	60.98	73.77	65.57	55.24	61.87
50%	Wanda	55.01	47.35	55.09	56.03	49.97	**75.40**	55.80	49.17	53.43
	SparseGPT	49.42	49.04	55.00	53.16	47.74	67.21	53.56	45.74	50.78
	ECoFLaP	51.33	**50.95**	52.19	52.61	49.88	63.93	54.52	46.01	51.47
	Multiflow	51.95	47.80	54.18	54.55	47.89	73.77	53.92	47.59	51.66
	PORTA	**55.23**	50.05	**55.81**	**56.50**	**51.46**	72.13	**56.82**	**49.76**	**54.30**

based approximations for one-shot pruning; ECoFLaP [31], a zeroth-order pruning method tailored for VLMs; and Multiflow [9], a task-agnostic pruning approach built upon modality-specific weight distributions. Following our zero-shot protocol, we report Multiflow results without any task-specific fine-tuning; additional comparisons with Multiflow fine-tuning are provided in Appendix A. For calibration, we estimate activation variation from a general image–text set. We use MSCOCO as a standard, publicly available calibration source, which contains 82,783 image–text pairs. Importantly, PORTA does not rely on large-scale or downstream-aligned calibration.

All experiments are conducted on a single NVIDIA L40S GPU. We focus on the standard one-shot pruning setting and evaluate sparsity levels of 45%, 50%, 55%, 60%, and 65%. To ensure a fair comparison across pruning methods, all evaluations use identical calibration data and identical sample counts, so that performance differences arise solely from the pruning strategies.

4.2 Main Experimental Results

We report zero-shot performance for CLIP models pruned by PORTA and baseline methods under high sparsity levels of 0.55–0.65 across retrieval and classification benchmarks in Tab. 3. Overall, PORTA maintains stronger performance than existing pruning baselines in high-sparsity regimes, with particularly consistent advantages over ECoFLaP, whose performance degrades markedly beyond 50% sparsity. Moreover, while LLM-oriented importance formulations used by Wanda and SparseGPT may not fully align with the characteristics of vision–

language representations, PORTA achieves the best or competitive results in the majority of settings, suggesting that its importance formulation is better aligned with VLM representations. Quantitatively, PORTA's relative advantage tends to become more visible as sparsity increases. At 65% sparsity, PORTA improves the mean performance over the eight reported metrics by 12.6% compared to SparseGPT and by 21.5% compared to Wanda, as shown in Tab. 3. For example, PORTA improves MSCOCO IR@5 from 35.00 to 37.32 over SparseGPT and maintains higher ImageNet accuracy at 31.38% versus 25.96%. On BLIP at 45% sparsity, as reported in Tab. 4, PORTA attains a higher mean score of 90.26 than 89.95 for SparseGPT and improves TR@1 from 81.70 to 83.00, indicating reliable effectiveness beyond CLIP.

We further validate PORTA beyond CLIP/BLIP in generative VLM settings. On Qwen2-VL [36] for ScienceQA [25] at 50% sparsity, as shown in Tab. 5, PORTA achieves the highest average accuracy of 54.30 compared to 61.87 for the dense model. PORTA also outperforms all pruning baselines; in particular, it surpasses the second-best baseline Wanda by 0.87%p, increasing the average accuracy from 53.43 to 54.30. PORTA achieves the best average accuracy and remains competitive across categories, with particularly strong performance in natural science with 55.23, language with 55.81, and text-based context with 56.50. These results demonstrate that our importance formulation transfers effectively to generative VLMs and visual QA tasks. Additional results on diffusion-based models and OpenFlamingo [3] image captioning [4] are provided in Appendix B and C, respectively.

4.3 Pruning Ratio Visualization

We visualize the layer-wise pruning ratios assigned by PORTA when pruning CLIP to a 60% target sparsity in Fig. 3. For visualization, we use a fixed α; additional analysis is provided in Appendix D. At the modality level, the language encoder exhibits relatively uniform pruning ratios in the range of 55%–65%, whereas the vision encoder shows substantially larger variation, spanning approximately 45%–65%. Across both encoders, early layers tend to receive higher sparsity, while deeper layers are assigned lower sparsity, reflecting their greater contribution to downstream representations—an observation consistent with prior studies on representation depth in VLMs. Previous work [11] has shown that the final layers of CLIP models are responsible for producing the majority of its semantic representations, and [8] further demon-

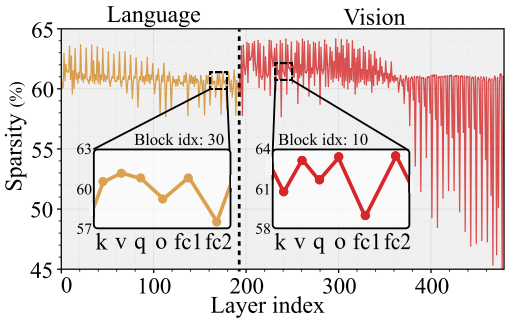

Fig. 3: Layer-wise pruning ratio allocation produced by PORTA on CLIP at 60% sparsity ($\alpha = 10$).

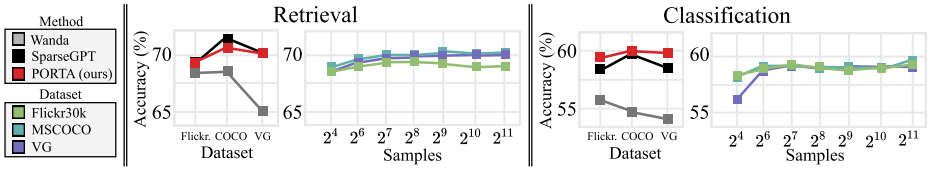

Fig. 4: Effect of calibration data type, pruning method, and sample size on zero-shot performance at 60% sparsity with CLIP.

strates that deeper layers exhibit increasing representational complexity and encode progressively more diverse concepts.

4.4 Calibration Data and Sample Size

We evaluate the calibration robustness of PORTA on CLIP at 60% sparsity under zero-shot retrieval and classification in Fig. 4. The left plot varies the calibration distribution across Flickr30K, MSCOCO, and Visual Genome [17]. While prior methods show noticeable performance shifts as the calibration source changes, with accuracy fluctuations reaching up to about 5%, PORTA remains substantially more stable, staying within roughly 1% across datasets. This stability is consistently observed in both retrieval and classification, indicating that PORTA is robust to the choice of calibration source.

The right plot varies the number of calibration samples from 2^4 to 2^{11}. Except for the smallest setting, PORTA exhibits near-flat performance across a wide range of sample counts, indicating that it does not require large-scale calibration to obtain reliable importance estimates. Together, these results support a practical prune-once workflow: PORTA can be calibrated with a general-purpose image–text set and reused across heterogeneous downstream tasks without task-specific recalibration or repruning, thereby reducing deployment overhead. Further robustness results on additional calibration datasets are provided in Appendix E.

4.5 Ablation between Importance and Sparsity Allocation

To analyze the respective contributions of the importance score and the sparsity allocation strategy, we conduct a cross-combination study under 60% sparsity, pairing the magnitude-based importance score used by Wanda with the proposed variance-based PORTA score. As shown in Tab. 6, replacing the magnitude-based Wanda score with the

Table 6: Cross-combination study of score and sparsity-allocation strategies at 60% sparsity on CLIP. Combining the PORTA score with the PORTA ratio yields the best performance across all benchmarks.

Sparsity	Score	Ratio	CIFAR10	CIFAR100	ImageNet
60%	Wanda	×	93.20	70.28	57.35
	PORTA	×	94.91	74.45	59.35
	Wanda	Multiflow	10.03	1.58	0.11
	PORTA	Multiflow	10.07	1.40	0.11
	Wanda	ECoFLaP	93.52	71.16	61.14
	PORTA	ECoFLaP	93.57	71.19	61.33
	Wanda	**PORTA**	94.62	74.29	59.04
	PORTA	**PORTA**	**95.06**	**75.59**	**59.66**

Table 7: Ablation of the zero-mean assumption and diagonal covariance approximation at 60% sparsity on CLIP retrieval.

Method	TR@1	TR@5	IR@1	IR@5	Avg.
PORTA	60.14	83.04	43.11	68.18	63.62
PORTA w/o zero-mean	60.20	83.12	43.11	68.12	63.64
PORTA w/o diagonal approximation	60.70	82.80	43.13	68.27	63.73

proposed variance-based PORTA score consistently improves performance across datasets, even under the same sparsity allocation scheme. Notably, when sparsity is uniformly distributed across layers, PORTA improves CIFAR-100 accuracy from 70.28 to 74.45 and ImageNet accuracy from 57.35 to 59.35, indicating that variance provides a stronger and more stable pruning signal in multimodal settings. When comparing sparsity allocation strategies under the same importance score, the proposed PORTA ratio consistently improves performance over uniform allocation and remains competitive or superior to ECoFLaP across datasets.

Under the PORTA score, replacing uniform allocation with the proposed PORTA ratio improves CIFAR-100 accuracy from 74.45 to 75.59 and ImageNet accuracy from 59.35 to 59.66. Similarly, under the magnitude-based score, the PORTA ratio improves CIFAR-100 from 70.28 to 74.29. Notably, combining the PORTA score with the PORTA ratio achieves the best overall performance, suggesting that variance-based importance estimation and output-variance–guided sparsity allocation complement each other. These results confirm that both components contribute to the final gains of PORTA, and that the improvement cannot be attributed to either factor alone.

4.6 Ablation on Modeling Assumptions

To examine the effect of the zero-mean assumption and the diagonal covariance approximation used in Sec. 3.3, we conduct an ablation study under 60% sparsity. As shown in Tab. 7, removing the zero-mean assumption results in negligible performance differences, by at most 0.1%. Eliminating the diagonal approximation leads to at most a 1% fluctuation. These results indicate that both assumptions have minimal impact on retrieval performance.

Given that the diagonal approximation enables efficient computation and allows reuse of variance statistics from the importance estimation stage for pruning-ratio calculation, we adopt the diagonal approximation in PORTA.

4.7 Computational Cost Analysis

One practical concern in one-shot pruning is the additional computational overhead introduced by ratio allocation across layers. To evaluate the practical efficiency of PORTA, we compare the wall-clock pruning time with representative one-shot pruning baselines under the same calibration setting.

Table 8: Pruning time comparison on CLIP. Best result is boldfaced and second-best is underlined.

	PORTA	**SparseGPT**	**Wanda**	**ECoFLaP**	**Multiflow**
Pruning Time (s)	<u>195.34</u>	773.35	**151.61**	402.28	229.65

As shown in Tab. 8, PORTA completes pruning in ∼195s, including layer-wise ratio allocation, achieving 3.96× and 2.06× faster than SparseGPT and ECoFLaP, respectively. This efficiency is enabled by reusing the variance statistics computed for importance estimation when calculating pruning ratios via the diagonal covariance approximation discussed in Sec. 4.6.

5 Conclusion

We presented PORTA, a prune-once, retraining-free framework for task-agnostic compression of vision–language models. PORTA is motivated by the observation that commonly used magnitude-based importance measures can be strongly influenced by modality-specific activation characteristics, which may introduce modality-sensitive bias and lead to imbalanced pruning across the vision and language branches. To obtain a more reliable indicator of representation utility, PORTA integrates activation variance statistics into importance estimation, yielding a pruning criterion that is less sensitive to modality-dependent scale and distribution effects. In addition, we introduce a strategy for layer-wise sparsity allocation guided by output feature variance, improving compression robustness compared with existing pruning methods under high sparsity levels. Extensive zero-shot evaluations across representative VLM families and diverse task categories demonstrate that PORTA consistently outperforms strong pruning baselines at the same sparsity levels, while remaining stable under changes in calibration data and sample size. Overall, PORTA advances practical VLM deployment by enabling a single pruned model to generalize across heterogeneous downstream applications without re-pruning or retraining.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant (RS-2025-00555943); by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants (No. RS-2021-II211341; Artificial Intelligence Graduate School Program (Chung-Ang University), RS-2026-25513331; Digital Columbus Project, IITP-2026-RS-2026-25546026 and IITP-2024-RS-2024-00397085; Leading Generative AI Human Resources Development, RS-2025-02219317; AI Star Fellowship (Kookmin University), and IITP-2024-RS-2024-00417958; Global Research Support Program in the Digital Field) funded by the Korea government (MSIT).

Bibliography

[1] Alizadeh, M., Tailor, S.A., Zintgraf, L.M., van Amersfoort, J., Farquhar, S., Lane, N.D., Gal, Y.: Prospect pruning: Finding trainable weights at initialization using meta-gradients. In: International Conference on Learning Representations (ICLR) (2022)

[2] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. Proceedings of the IEEE international conference on computer vision. (2015)

[3] Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., Jitsev, J., Kornblith, S., Koh, P.W., Ilharco, G., Wortsman, M., Schmidt, L.: OpenFlamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023)

[4] Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., Plank, B.: Automatic description generation from images: A survey of models, datasets, and evaluation measures. Journal of Artificial Intelligence Research (JAIR) **55**, 409–442 (2016)

[5] Blalock, D., Ortiz, J.J.G., Frankle, J., Guttag, J.: What is the state of neural network pruning? In: Proceedings of Machine Learning and Systems (MLSys) (2020)

[6] Cao, M., Li, S., Li, J., Nie, L., Zhang, M.: Image-text retrieval: A survey on recent research and development. arXiv preprint arXiv:2203.14713 (2022)

[7] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)

[8] Dorszewski, T., Tětková, L., Jenssen, R., Hansen, L.K., Wickstrøm, K.K.: From colors to classes: Emergence of concepts in vision transformers. In: Explainable Artificial Intelligence, pp. 28–47. Springer (2025)

[9] Farina, M., Mancini, M., Cunegatti, E., Liu, G., Iacca, G., Ricci, E.: Multiflow: Shifting towards task-agnostic vision-language pruning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16185–16195 (2024)

[10] Frantar, E., Alistarh, D.: SparseGPT: Massive Language Models Can Be Accurately Pruned in One-Shot. In: ICML (2023)

[11] Gandelsman, Y., Efros, A.A., Steinhardt, J.: Interpreting clip's image representation via text-based decomposition. arXiv preprint arXiv:2310.05916 (2023)

[12] Han, S., Mao, H., Dally, W.J.: Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. NeurIPS (2015)

[13] Han, S., Pool, J., Tran, J., Dally, W.: Learning both weights and connections for efficient neural network. NeurIPS (2015)

[14] He, X., Sygnowski, J., Galashov, A., Rusu, A.A., Teh, Y.W., Pascanu, R.: Task agnostic continual learning via meta learning. In: 4th Lifelong Machine Learning Workshop at ICML (2020)

[15] Jamal, M.A., Qi, G.J.: Task agnostic meta-learning for few-shot learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11719–11727 (2019)

[16] de Jorge, P., Sanyal, A., Behl, H.S., Torr, P.H.S., Rogez, G., Dokania, P.K.: Progressive skeletonization: Trimming more fat from a network at initialization. In: International Conference on Learning Representations (ICLR) (2021)
[17] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1), 32–73 (2017)
[18] Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep., University of Toronto (2009)
[19] Lee, J., Park, S., Mo, S., Ahn, S., Shin, J.: Layer-adaptive Sparsity for the Magnitude-based Pruning. In: ICLR (2021)
[20] Lee, N., Ajanthan, T., Torr, P.: SNIP: Single-Shot Network Pruning based on connection sensitivity. In: ICLR (2019)
[21] Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
[22] Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022)
[23] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: ECCV. Springer (2014)
[24] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
[25] Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. In: NeurIPS (2022)
[26] Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Indian Conference on Computer Vision, Graphics & Image Processing (2008)
[27] Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015)
[28] 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: ICML (2021)
[29] Shi, D., Tao, C., Jin, Y., Yang, Z., Yuan, C., Wang, J.: UPop: Unified and progressive pruning for compressing vision-language transformers. In: ICML (2023)
[30] Sun, M., Liu, Z., Bair, A., Kolter, J.Z.: A Simple and Effective Pruning Approach for Large Language Models. ICLR (2024)
[31] Sung, Y.L., Yoon, J., Bansal, M.: Ecoflap: Efficient coarse-to-fine layer-wise pruning for vision-language models. In: International Conference on Learning Representations (ICLR) (2024)
[32] Tanaka, H., Kunin, D., Yamins, D.L., Ganguli, S.: Pruning neural networks without any data by iteratively conserving synaptic flow. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
[33] Wang, C., Zhang, G., Grosse, R.: Picking winning tickets before training by preserving gradient flow. In: International Conference on Learning Representations (ICLR) (2020)
[34] Wang, H., Zhang, J., Ma, Q.: Exploring intrinsic dimension for vision-language model pruning. In: Forty-first International Conference on Machine Learning (2024)

[35] Wang, H., Wang, S., Zhang, W.Q., Suo, H., Wan, Y.: Task-agnostic structured pruning of speech representation models. In: Proc. Interspeech 2023. pp. 231–235 (2023)
[36] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
[37] Wang, T., Zhou, W., Zeng, Y., Zhang, X.: EfficientVLM: Fast and accurate vision-language models via knowledge distillation and modal-adaptive pruning. In: Findings of the Association for Computational Linguistics: ACL 2023 (2023)
[38] Wang, Y., Li, D., Sun, R.: Ntk-sap: Improving neural network pruning by aligning training dynamics. In: International Conference on Learning Representations (ICLR) (2023)
[39] Williams, M., Aletras, N.: How does calibration data affect the post-training pruning and quantization of large language models? arXiv preprint arXiv:2311.09755 (2023)
[40] Xu, J., Tan, X., Luo, R., Song, K., Li, J., Qin, T., Liu, T.Y.: Nas-bert: Task-agnostic and adaptive-size bert compression with neural architecture search. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. pp. 1933–1943 (2021)
[41] Yin, L., Wu, Y., Zhang, Z., Hsieh, C.Y., Wang, Y., Jia, Y., Li, G., Jaiswal, A., Pechenizkiy, M., Liang, Y., et al.: Outlier weighed layerwise sparsity (owl): A missing secret sauce for pruning llms to high sparsity. arXiv preprint arXiv:2310.05175 (2023)
[42] Zhang, Z., Gong, B., Chen, Y., Han, X., Zeng, G., Zhao, W., Chen, Y., Liu, Z., Sun, M.: Bmcook: A task-agnostic compression toolkit for big models. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 396–405 (2022)