Unified Knowledge Maintenance Pruning and Progressive Recovery with Weight Recalling for Large Vision-Language Models

Zimeng Wu[1,2], **Jiaxin Chen**[1,2*], **Yunhong Wang**[1,2]

[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[2]School of Computer Science and Engineering, Beihang University, Beijing, China
{zimengwu, jiaxinchen, yhwang}@buaa.edu.cn

Abstract

Large Vision-Language Model (LVLM), leveraging Large Language Model (LLM) as the cognitive core, has recently become one of the most representative multimodal model paradigms. However, with the expansion of unimodal branches, *i.e.* visual encoder and LLM, the storage and computational burdens intensify, posing challenges for deployment. Structured pruning has proved promising in compressing large models by trimming a large portion of insignificant network structures. Nevertheless, most of them are predominantly designed for LLMs, either relying on unitary importance metrics that fail to deal with modality-wise imbalances or adopting generic pruning and recovery paradigms that overlook the unique calibration status and capability requirements of large models, leading to substantial performance degradation. To address these issues, we propose a novel structured pruning approach for LVLMs, dubbed Unified Knowledge Maintenance Pruning and Progressive Recovery with Weight Recalling (UKMP). Specifically, we design a Unified Knowledge Maintenance Importance (UKMI) metric, which simultaneously considers balancing the block-wise and modality-wise importance by adaptive normalization, optimizing the importance estimation by refining gradient-based criteria, and maintaining the knowledge capacity of LVLMs by using the angle distribution information entropy. Moreover, we develop a LoRA-based Progressive Distillation (LPD) method that recalls the pruned weights and performs progressive distillation for comprehensive recovery. Extensive experimental results across various vision-language tasks demonstrate the effectiveness of our approach, comparing to the state-of-the-art structured pruning methods.

Code — https://github.com/Wuzimeng/UKMP.git

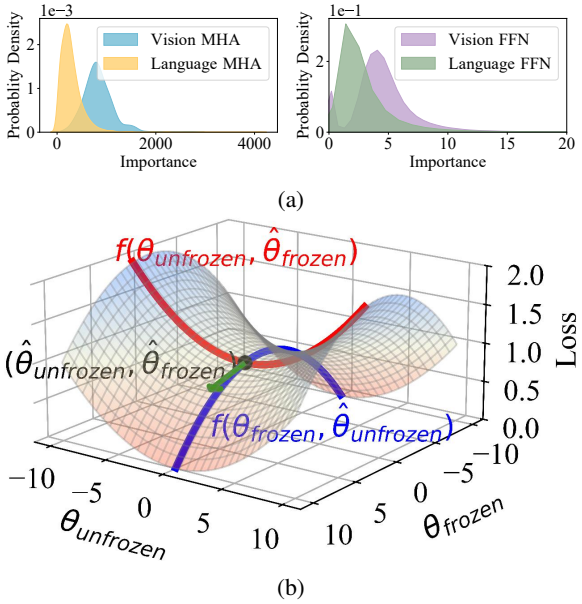

Figure 1: Main challenges in pruning LVLMs. (a) Existing works generate severely imbalanced parameter importance for distinct modalities and network modules. (b) The pre-training of LVLMs mostly optimizes a part of parameters for efficiency with the rest frozen before pruning, resulting in unconvergence of the frozen parameters (*i.e.* unfrozen parameters converge to the black dot along the red line, but the model can still improve along the blue line), which violates the convergence assumption of existing pruning methods.

Introduction

Large Vision-Language Models (LVLMs), which are built upon Large Language Models (LLMs) by aligning visual features to the frozen language feature space, have recently emerged as a prevalent paradigm (Dai et al. 2023; Alayrac et al. 2022). Leveraging their powerful language generation and zero-shot learning capabilities, LVLMs deliver exceptional performance on various tasks such as visual question answering, image-text retrieval and image captioning (Li et al. 2023b). However, the ever-expending scale of models for both vision and language modalities inevitably increases computational demands and inference latency, limiting their deployment on resource-constrained platforms.

In order to reduce model size and computational complexity, various model compression techniques have been explored for adaptating large models, including quantization (Frantar et al. 2023; Wu et al. 2025), token pruning (Cao et al. 2024), and more. Structured network pruning, offering flexibility in scaling without relying on hardware-specific optimizations, emerges as a promising solution for

*Corresponding Author

explicit acceleration. Unlike pre-LLM era strategies, pruning for large models is generally task-agnostic (Sung, Yoon, and Bansal 2024). Given the vast data and parameter scales involved, most methods estimate parameter importance on a small-scale calibration dataset, followed by one-shot pruning and resource-efficient recovery (Ma, Fang, and Wang 2023; An et al. 2024). However, existing methods are primarily designed for either standalone LLMs (Ashkboos et al. 2024) or the LLM component within LVLMs (Wang et al. 2024), yielding suboptimal performance when applied to unified pruning of LVLMs.

This limitation mainly arises from three factors: 1) Absence of global imbalance handling within structured pruning schemes. Although some works have explored globally adaptive sparsity in LVLMs (Sung, Yoon, and Bansal 2024; He, Li, and Chen 2024), these efforts are restricted to unstructured schemes and do not address the module-wise and modality-wise imbalances inherent in structured pruning, as shown in Fig. 1(a). 2) Overlooking changes in calibration status of large models by simply using importance criteria for traditional models. As displayed in Fig. 1(b), unimodal components, typically frozen during LVLM pre-training, are unconvergent prior to pruning, and external calibration data may not align with the pre-training data. This mismatch results in suboptimal weight importance estimation when using widely adopted gradient-based criteria (Ma, Fang, and Wang 2023). 3) Neglecting maintaining the capability of large models, when applying existing singular criteria and recovery techniques (Kim et al. 2024). Concretely, most methods fail to preserve the internal knowledge that underpins the remarkable zero-shot learning capability of large models. This issue is more pronounced in LVLMs, when the vision-language pre-training primarily focuses on modality alignment (Yin et al. 2024), while the zero-shot performance heavily depends on external knowledge embedded in the frozen LLM, leading to performance degradation.

To address these issues, we propose a novel structured pruning approach for LVLMs, dubbed Unified Knowledge Maintenance Pruning and Progressive Recovery with Weight Recalling (UKMP). We firstly develop a Unified Knowledge Maintenance Importance metric (UKMI), which handles the imbalance in unified pruning by using adaptive normalization across blocks and modalities based on the one-shot estimated importance metric. The Taylor importance metric is refined by selecting poorly fitted token prediction sub-tasks during backpropagation. Subsequently, to preserve the internal knowledge of LVLMs, we perceive the correlation between parameters and model's knowledge capacity by angular distribution entropy, and integrate it into the gradient-based importance. For efficient recovery after pruning, we further design the LoRA-based Progressive Distillation (LPD). This module recalls pruned weights to reuse the inherit knowledge and employs progressive learning to minimize excessive compensation between modalities.

Our main contributions are summarized in three-fold:

- We propose a novel pruning approach, dubbed UKMP, for large vision-language models. To the best of our knowledge, we make the first attempt to explore unified structured pruning for LVLMs.
- We develop a unified knowledge maintenance importance metric and a LoRA-based progressive distillation process to address modality imbalance, error in gradient-based criteria, and loss of knowledge during pruning.
- We conduct extensive experiments across various vision-language tasks for evaluation. The results demonstrate that our method significantly outperforms existing state-of-the-art methods, especially at high pruning ratios.

Related Work

Large Vision-Language Models

Large Vision-Language Models (LVLMs) are advanced multimodal models that leverage capabilities of LLMs to integrate vision and language for complex tasks (Liu et al. 2024) such as image captioning (Li et al. 2023b), VQA (Dai et al. 2023), video understanding (Team et al. 2024), etc. Typically, LVLMs consist of a pre-trained visual encoder, a pre-trained LLM and a modality interface (Yin et al. 2024). In prevalent architectures, both modalities are transformer-based and the modality interface is lightweight (Li et al. 2023b). As for the pre-training, it is common practice to freeze modality-specific modules and optimize the interface for modality alignment (Pi et al. 2023). In this paper, we focus on the compression of LVLMs trained in this manner to develop task-agnostic sub-models.

Network Pruning

General Pruning Methods Network pruning simultaneously reduces memory size and improves inference speed (Liang et al. 2021), and is typically categorized into unstructured and structured approaches. While unstructured and intermediate semi-structured pruning create sparse matrices that require specialized hardware for acceleration (Zhang et al. 2022; Fang, Zhou, and Wang 2022), structured pruning directly reduces matrix sizes, enabling flexible deployment (Fang et al. 2023). Network pruning typically involves estimating parameter importance, removing redundancies, and retraining for recovery (Lin et al. 2024; Shi et al. 2023). Gradient-based criteria, especially those using Taylor expansion to approximate parameter-loss correlation, demonstrate effectiveness across models, and have extended from unstructured to structured pruning (LeCun, Denker, and Solla 1989; Jiang et al. 2023; Wu, Chen, and Wang 2023).

Pruning for LLMs and LVLMs Pruning for LLMs and LVLMs, given their vast parameter scales, emphasizes efficient, low-resource compression (Jin et al. 2024; Frantar and Alistarh 2023). Importance metrics are commonly derived via gradient backpropogation on small datasets (Ma, Fang, and Wang 2023) or solely by forward passing (Ashkboos et al. 2024; Sun et al. 2023a). Structured pruning often involves post-training recovery, such as low-rank approximation (LoRA) (Hu et al. 2021), under limited data (Ma, Fang, and Wang 2023), while some methods use training-free reconstruction (An et al. 2024). Although methods like ECoFLaP (Sung, Yoon, and Bansal 2024) and SparseLoRA (He, Li, and Chen 2024) have explored pruning for LVLMs, these efforts are limited to unstructured schemes. AntGroup's structured pruning for AntGMM

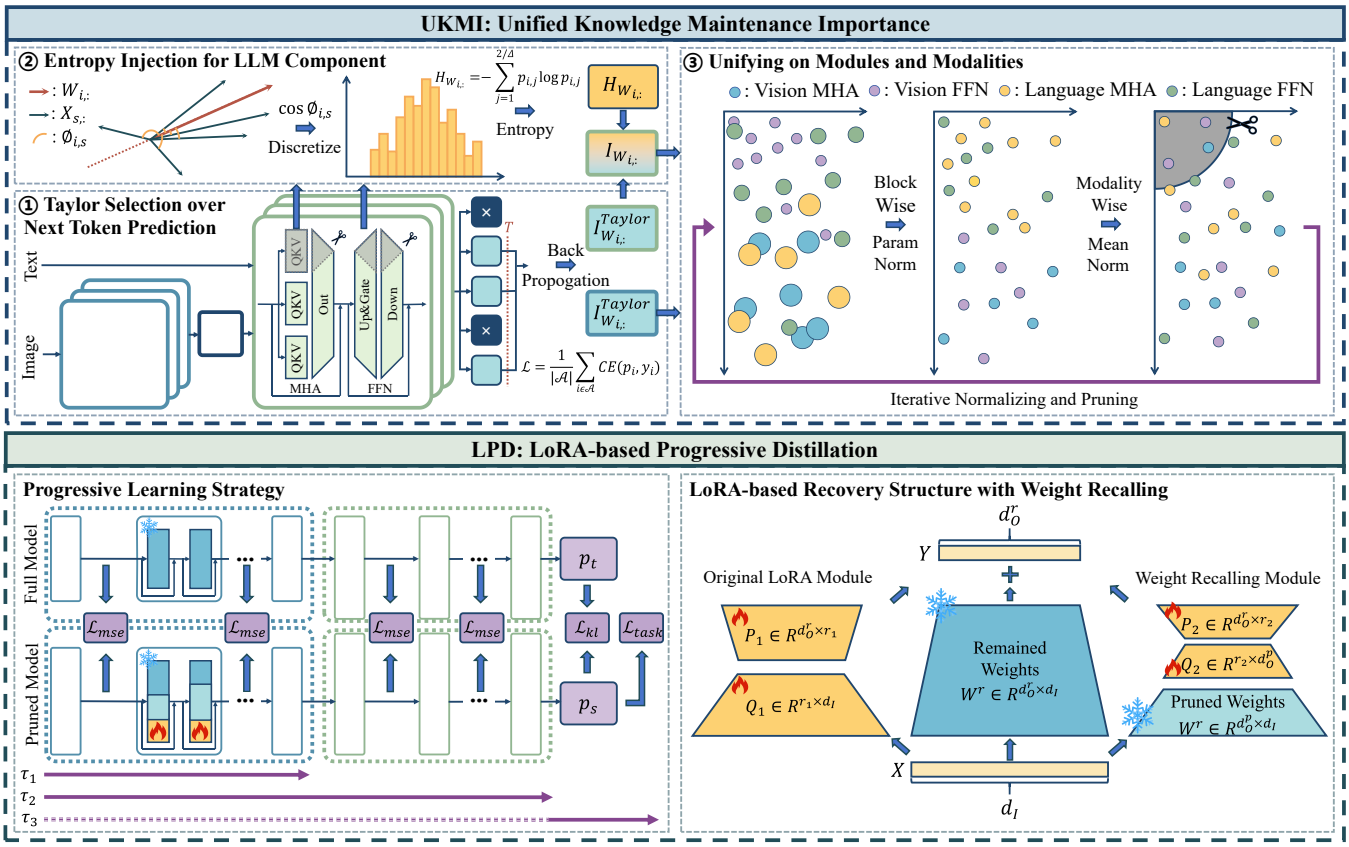

Figure 2: Framework of our proposed UKMP. During pruning, the UKMI module sequentially performs: ① Select the poorly fitted token prediction sub-tasks to compose loss and backpropagate for revised Taylor importance. ② Calculate the information entropy of angle distribution between token and weight vectors, and integrate it to the Taylor importance. ③ Iteratively normalize the pre-computed importance and prune till target pruning ratio. During recovery, the LPD module injects a LoRA-based structure with weight recalling to each weight matrix and then employs a progressive learning strategy for distillation.

(Wang et al. 2024) targets only the LLM component and classification tasks. In contrast, structured pruning receives more attention on LLMs, yet despite the structural similarities, these methods often underperform on LVLMs, especially at high pruning ratios. Therefore, structured pruning for LVLMs remains an underexplored but valuable area.

Methodology

Framework Overview

Pruning is the process of searching a sub-model that has minimal impact to the full model. Take the loss function $\mathcal{L}$ as a reflection of model performance and given the target pruning ratio $\mathcal{P}$, pruning can be cast as an optimization problem:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\hat{\boldsymbol{\theta}} \in \boldsymbol{\theta}} \mathcal{L}(\mathcal{D}; \hat{\boldsymbol{\theta}}), \text{ s.t.} |\hat{\boldsymbol{\theta}}| \leq |\boldsymbol{\theta}| \cdot (1 - \mathcal{P}), \quad (1)$$

where $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ are the set of weights from the full model and the pruned model, respectively. $\mathcal{D}$ is the calibration dataset. To reduce the complexity of the search process, gradient-based criterion defines the parameter importance as the loss change when discarding it. Then for a parameter $\boldsymbol{\theta}_i$, applying Taylor expansion yields the Taylor importance:

$$\begin{aligned} I_{\boldsymbol{\theta}_i}^{Taylor} &= |\Delta\mathcal{L}(\mathcal{D})| = |\mathcal{L}_{\boldsymbol{\theta}_i}(\mathcal{D}) - \mathcal{L}_{\boldsymbol{\theta}_i=0}(\mathcal{D})| \\ &= |\frac{\partial \mathcal{L}(\mathcal{D})}{\partial \boldsymbol{\theta}_i}\boldsymbol{\theta}_i - \frac{1}{2}\boldsymbol{\theta}_i \boldsymbol{H}_{ii} \boldsymbol{\theta}_i + O(||\boldsymbol{\theta}_i||^3)| \\ &\approx |\frac{\partial \mathcal{L}(\mathcal{D})}{\partial \boldsymbol{\theta}_i}\boldsymbol{\theta}_i - \frac{1}{2}(\frac{\partial \mathcal{L}(\mathcal{D})}{\partial \boldsymbol{\theta}_i}\boldsymbol{\theta}_i)^2|, \end{aligned} \quad (2)$$

where typically, the third-order infinitesimal terms are discarded and the second term is approximated by the Fisher Information theory to avoid quadratic computational complexity of the hessian matrix $\boldsymbol{H}$.

In context of structured pruning, given a weight matrix $\boldsymbol{W} \in \mathbb{R}^{d_O \times d_I}$, the goal is to search a binary mask $\boldsymbol{M}$ of length d_O or d_I that performs Hadamard product with $\boldsymbol{W}$ as $\hat{\boldsymbol{W}} = \boldsymbol{W} \odot \boldsymbol{M}$, indicating which rows or columns to prune. With $I_{\boldsymbol{W}_{i,j}}$ representing the importance of a single parameter, the importance of a row $\boldsymbol{W}_{i,:}$ is usually defined as the summation of all parameter importance within it, as shown in Eq.(3), and similarly for columns. Specifically, $I_{\boldsymbol{W}_{i,j}}$ can

be derived from Eq.(2) if the Taylor importance is employed.

$$I_{\boldsymbol{W}_{i,:}} = \sum_{j=1}^{d_I} I_{\boldsymbol{W}_{i,j}}. \quad (3)$$

We explore pruning for popular transformer-based LVLMs, where both the visual encoder and LLM consist of stacked Multi-Head Attention (MHA) blocks and Feed-Forward Network (FFN) blocks. Due to the coupled structures, multiple vectors need to be grouped, with corresponding rows or columns pruned together (Fang et al. 2023). Thus, Eq.(3) is expanded to the sum of parameter importance across the entire group:

$$I_{\boldsymbol{W}_{i,:}} = I_{\mathcal{G}} = \sum_{\boldsymbol{W}_{k,:}\in\mathcal{G}} \sum_{j=1}^{d_I} I_{\boldsymbol{W}_{k,j}} + \sum_{\boldsymbol{W}_{:,k}\in\mathcal{G}} \sum_{j=1}^{d_O} I_{\boldsymbol{W}_{j,k}}, \quad (4)$$

where $\mathcal{G}$ is the coupled group satisfying $\boldsymbol{W}_{i,:} \in \mathcal{G}$. We prune the inter-block dimensions for finer granularity. Notably, for the parallel computation on regular devices, vectors are grouped by attention heads in MHA blocks.

As for the overall process, we follow the typical paradigm of pruning large models, *i.e.* groups are sequentially removed according to their importance order till the target pruning ratio, followed by an efficient training for recovery. Furthermore, to tackle the aforementioned challenges in unified pruning for LVLMs, we propose the UKMP, by developing a Unified Knowledge Maintenance Importance metric (UKMI) for pruning and a LoRA-based Progressive Distillation process (LPD) for recovery, as shown in Fig. 2. For clarity, we provide detailed description in a logical sequence that differs from the actual execution order.

Unified Knowledge Maintenance Importance

Unifying on Modules and Modalities In the literature, unified pruning is prone to incur imbalances in the importance metric across structures, leading to over-pruning. Similar issues also arise in LVLMs, where imbalances occur in both distinct modules and modalities as in Fig. 1(a).

According to Eq.(4), the accumulation operatioin within groups is one of the main causes. Thus, we perform block-wise normalization by parameter volume within groups, pruning the group with the lowest average importance:

$$I^{bn}_{\boldsymbol{W}_{i,:}} = I^{bn}_{\mathcal{G}} = I_{\boldsymbol{W}_{i,:}}/\mathrm{Param}(\mathcal{G}) = I_{\boldsymbol{W}_{i,:}}/(d_I \cdot |\mathcal{G}|), \quad (5)$$

where $\boldsymbol{W}_{i,:} \in \mathcal{G}$ and $|\mathcal{G}|$ represents the number of coupled elements. Then, for the difference between modalities, we empirically scale their distributions to a similar range using mean normalization:

$$I^{mn}_{\boldsymbol{W}^v_{i,:}} = I^{mn}_{\mathcal{G}} = \frac{I^{bn}_{\boldsymbol{W}^v_{i,:}} \cdot |\mathcal{G}_V|}{\sum_{\mathcal{G}_v\in\mathcal{G}_V} I^{bn}_{\mathcal{G}_v}}. \quad (6)$$

where $\mathcal{G}_V$ denotes the set of coupled groups in the vision component, and similarly for the LLM component. In order to capture distribution changes in a more fine-grained manner, we iteratively normalize and prune. Notably, the importance metric is computed before the iterations, thus the iterative process remains computationally efficient.

Taylor Selection over Next Token Prediction Due to the unique calibration status of LVLMs, the Taylor importance needs revision. For small models, the first term of Eq.(2) is often neglected since all parameters converge on $\mathcal{D}$. While for LVLMs, this term remains significant, as the pre-trained model retains a first-order gradient in its frozen parameters (see Fig. 1(b)). Meanwhile, the second term has been experimentally shown by prior researches to introduce fluctuations in results (Ma, Fang, and Wang 2023). We attribute this to the limitations of Fisher Information approximation, which requires $\mathcal{D}$ to match the model's distribution. However, external sourcing or conditional filtering of $\mathcal{D}$ tends to change data distribution, leading to an increase in error (detailed discussion is provided in *Appendix*). Therefore, we conclude that only the first term should be utilized for parameter importance estimation in LVLMs, formally as:

$$I^{Taylor}_{\boldsymbol{W}_{i,j}} \approx |\frac{\partial \mathcal{L}(\mathcal{D})}{\partial \boldsymbol{W}_{i,j}} \boldsymbol{W}_{i,j}|. \quad (7)$$

Furthermore, we regard the training process with a single image-text pair as a combination of multiple independent next-token prediction sub-tasks. Inspired by (Fang, Ma, and Wang 2023), when retaining only the first term of Taylor importance, sub-tasks contribute differently to the optimization objective. For sub-tasks that the network fits well, the first term tends to introduce noise, thereby interfering with importance estimation. To address this, we set a simple but effective hard threshold $\mathcal{T}$ to the predicted probability p_i of the ground truth token y_i to filter the sub-tasks, formally as Eq.(8), where $\mathrm{CE}(p_i, y_i)$ is the typically used cross-entropy loss and $\mathcal{A}$ denotes the set of maintained sub-tasks.

$$\mathcal{L} = \frac{1}{|\mathcal{A}|} \sum_{i\in\mathcal{A}} \mathrm{CE}(p_i, y_i), \mathcal{A} = \{i | p_i < \mathcal{T}\}. \quad (8)$$

Entropy Injection for LLM Component The gradient-based criterion reflects the impact of model parameters on the loss function, solely representing the modality alignment or language generation abilities. However, it overlooks the broad knowledge within LVLMs, which is crucial for their remarkable zero-shot performance. Given the limitations of the calibration dataset in covering all the learned knowledge, we instead observe the correlation between parameters and knowledge capacity with information entropy. Denoting the input sequence with length S as $\boldsymbol{X} \in \mathbb{R}^{S\times d_I}$, the interaction between input and weight matrix $\boldsymbol{W}$, formally as $\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{X}^T$, can be viewed as the projection of token vectors onto weight vectors. Excluding data scale effects, a higher dispersion in the angles between token and weight vectors suggests a broader range of knowledge handled by the weight vector. Therefore, we capture this dispersion and inject it into the importance metric exclusively for LLM.

Specifically, for a weight vector $\boldsymbol{W}_{i,:}$, let $\boldsymbol{\phi}_{i,s}$ represent the angle between it and the s-th token vector, we begin by calculating the cosine value:

$$\cos\phi_{i,s} = \frac{\boldsymbol{W}_{i,:} \cdot \boldsymbol{X}^T_{s,:}}{||\boldsymbol{W}_{i,:}||^2_2 \cdot ||\boldsymbol{X}_{s,:}||^2_2}, \quad (9)$$

Then, we discretize the range of cosine function (*i.e.* $[-1, 1]$) into $2/\Delta$ uniform bins, each bin ranging $b_j = [-1 + (j -$

$1)\Delta, -1+j\Delta)$. The probability of cosine values falling into each bin $p_{i,j}$ is calculated as $p_{i,j} = \frac{1}{S}\sum_{s=1}^{S} \mathbf{1}_{b_j}(\cos\phi_{i,s})$ and used to calculate the information entropy of angular distribution $H_{\boldsymbol{W}_{i,:}}$ as below:

$$H_{\boldsymbol{W}_{i,:}} = -\sum_{j=1}^{2/\Delta} p_{i,j}\log(p_{i,j}). \tag{10}$$

Here we fix the number of bins to 100, thus $\Delta = 0.02$. Finally, we weight the original Taylor importance metric with the entropy to obtain the final parameter importance:

$$I_{\boldsymbol{W}_{i,:}} = I^{Taylor}_{\boldsymbol{W}_{i,:}} \cdot H_{\boldsymbol{W}_{i,:}}. \tag{11}$$

LoRA-based Progressive Distillation

LoRA-based Recovery Structure with Weight Recalling Recovery training after pruning commonly initializes with the remaining parameters. However, pruned parameters, especially in the LLM component at high pruning ratios, may still contain valuable information learned from pre-training, which is difficult to replenish with insufficient image-text pairs. To address this gap, we introduce a weight recalling module to implicitly recall the residual knowledge.

The recalling is performed by applying a linear transformation to reintegrate the pruned parameters into the remaining ones. Building on the typical LoRA-based structure, it operates in parallel with the original LoRA module during training and can be reparameterized into weight updates. Given the low-rank nature of the overall weight updates (Hu et al. 2021), we similarly assume this linear transformation matrix with a low "intrinsic rank". Consider a weight matrix $\boldsymbol{W} \in \mathbb{R}^{d_O \times d_I}$, with $\boldsymbol{W}^r \in \mathbb{R}^{d^r_O \times d_I}$ and $\boldsymbol{W}^p \in \mathbb{R}^{d^p_O \times d_I}$ representing the remained and pruned weights, respectively. In the original LoRA module, the update $\Delta_1\boldsymbol{W}$ is decomposed as $\Delta_1\boldsymbol{W} = \boldsymbol{P}_1\boldsymbol{Q}_1$, where $\boldsymbol{P}_1 \in \mathbb{R}^{d^r_O \times r_1}$ and $\boldsymbol{Q}_1 \in \mathbb{R}^{r_1 \times d_I}$. In the weight recalling module, the update $\Delta_2\boldsymbol{W}$, is decomposed as $\Delta_2\boldsymbol{W} = \boldsymbol{P}_2\boldsymbol{Q}_2\boldsymbol{W}^p$, where $\boldsymbol{P}_2 \in \mathbb{R}^{d^r_O \times r_2}$ and $\boldsymbol{Q}_2 \in \mathbb{R}^{r_2 \times d^p_O}$. Thus the forward computation is:

$$\begin{aligned} f(\boldsymbol{X}) &= (\boldsymbol{W}^r + \Delta_1\boldsymbol{W} + \Delta_2\boldsymbol{W})\boldsymbol{X} + b \\ &= (\boldsymbol{W}^r\boldsymbol{X} + b) + (\boldsymbol{P}_1\boldsymbol{Q}_1)\boldsymbol{X} + (\boldsymbol{P}_2\boldsymbol{Q}_2\boldsymbol{W}^p)\boldsymbol{X}, \end{aligned} \tag{12}$$

where b is the bias from full model. During distillation process for recovery, only $\boldsymbol{P}_1$, $\boldsymbol{P}_2$, $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ are updated, thereby reducing training complexity.

Progressive Learning Strategy General end-to-end post-training or distillation methods may lead to excessive compensatory between modalities. Notably, internal features and probability distributions from the full model provide more effective supervision. Therefore, we propose a progressive learning strategy for distillation to enhance knowledge transfer under limited training data.

Our distillation training combines the pre-training loss $\mathcal{L}_{task}$, the mean squared error (MSE) loss $\mathcal{L}_{mse}$, and the Kullback-Leibler (KL) divergence loss $\mathcal{L}_{kl}$ as the optimization objective, with the full model serving as teacher and the pruned model as student. The process is divided into three phases in macro as Eq.(13): align hidden states of the vision modality at τ_1, align hidden states of both the vision and language modalities at τ_2, and use the pre-training loss and KL divergence loss for further distillation at τ_3.

$$\mathcal{L}_\tau = \begin{cases} \beta_1\mathcal{L}_{mse}(\boldsymbol{E}^v_s, \boldsymbol{E}^v_t), & \tau{=}\tau_1; \\ \beta_1\mathcal{L}_{mse}(\boldsymbol{E}^v_s, \boldsymbol{E}^v_t){+}\beta_2\mathcal{L}_{mse}(\boldsymbol{E}^l_s, \boldsymbol{E}^l_t), & \tau{=}\tau_2; \\ \mathcal{L}_{task}(y_s, y){+}\mathcal{L}_{kl}(p_s, p_t), & \tau{=}\tau_3. \end{cases} \tag{13}$$

where τ denotes the training phase, βs are hyperparameters, $\boldsymbol{E}$ represents hidden states, y, p are predicted labels and distributions, and s, t, v, l denote the student model, the teacher model, the vision component and the language component, respectively. Hidden states are normalized within the MSE loss to address the scale differences across N layers:

$$\mathcal{L}_{mse}(\boldsymbol{E}_s, \boldsymbol{E}_t) = \sum_{n=1}^{N} \left\| \frac{\boldsymbol{E}_{s,n}}{||\boldsymbol{E}_{s,n}||_2^2} - \frac{\boldsymbol{E}_{t,n}}{||\boldsymbol{E}_{t,n}||_2^2} \right\|_2^2. \tag{14}$$

Experimental Results and Analysis

Experimental Settings

Architecture By following (Sung, Yoon, and Bansal 2024), we evaluate our proposed UKMP on the encoder-decoder version of BLIP-2 (Li et al. 2023b), composed of ViT-g/14 from EVA-ViT (Sun et al. 2023b) and $\mathrm{FlanT5_{XL}}$ (Chung et al. 2024). Since the Q-Former constitutes only 2.7% of the total parameters, we freeze it and focus on compressing the visual encoder and the LLM branch.

Datasets and Evaluation Metrics Under the setting of task-agnostic compression, we evaluate the zero-shot capabilities of the compressed models on various datasets and tasks. Specifically, we report the Accuracy on VQAv2 (Goyal et al. 2017), OK-VQA (Marino et al. 2019), GQA (Hudson and Manning 2019) for the VQA task, provide top-1 Text Recall (TR@1) and top-1 image Recall (IR@1) on Flickr30k (Plummer et al. 2015) for image-text retrieval, and report CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015) and SPICE (Anderson et al. 2016) on NoCaps (Agrawal et al. 2019) for image captioning. We also report the macro average score (Macro Avg.) across these tasks.

Implementation Details For calibration and recovery, we use CC595K (Liu et al. 2023), filtered from CC3M (Sharma et al. 2018) and synthesized by BLIP (Li et al. 2022). For pruning, 1000 image-text pairs are randomly selected as the calibration set. As for recovery, the pruned model is trained on all 595K data for 1 epoch, with 40K for phase τ_1, 280K for phase τ_2 and 275K for phase τ_3 as in Eq. (13). All the experiments are conducted with LAVIS (Li et al. 2023a) and TP (Fang et al. 2023) on one NVIDIA A800 GPU. Detailed settings of hyperparameters are provided in *Appendix*.

Main Results on Vision-Language Tasks

We compare UKMP with representative state-of-the-art structured pruning methods, including LLM-Pruner (Ma, Fang, and Wang 2023), ECoFLaP (Sung, Yoon, and Bansal 2024), FLAP (An et al. 2024) and UPop (Shi et al. 2023). As these methods are not originally designed for LVLMs, we re-implement them using the open-source code. Note that

Method	Pruning Ratio	Visual Question Answering			Image Captioning		Image-Text Retrieval		Macro Avg.
		VQAv2	OK-VQA	GQA	NoCaps		Flickr30k		
			Accuracy		CIDEr	SPICE	TR@1	IR@1	
Full Model-g (3.9B)	-	63.11	41.18	43.97	102.62	13.42	92.9	85.62	63.26
Full Model-L (3.2B)	-	61.11	37.72	43.45	98.69	13.07	93.3	83.56	61.56
LLM-Pruner global	20%	59.15	36.07	43.35	96.96	12.90	95.6	86.44	61.50
LLM-Pruner local	20%	54.86	35.74	41.91	94.00	12.78	95.4	85.16	59.98
ECoFLaP_sp	20%	60.22	39.90	41.51	99.90	13.24	**95.9**	**87.14**	62.54
FLAP	20%	50.45	27.05	33.13	94.98	12.63	94.7	83.44	56.63
UPop	20%	49.64	24.95	35.49	93.71	12.54	94.8	83.50	56.38
UKMP (Ours)	20%	**61.79**	**41.28**	**44.01**	**101.97**	**13.39**	95.0	86.36	**63.40**
LLM-Pruner global	50%	18.53	9.77	12.01	62.86	10.57	75.1	60.02	35.55
LLM-Pruner local	50%	36.40	10.50	25.15	66.72	10.73	78.8	63.12	41.63
ECoFLaP_sp	50%	6.40	7.76	4.25	85.51	11.93	88.3	73.56	39.67
FLAP	50%	5.78	8.41	3.80	81.70	11.65	89.1	73.88	39.19
UPop	50%	3.16	5.12	3.45	90.90	12.42	93.4	79.82	41.18
UKMP (Ours)	50%	**47.81**	**27.38**	**35.15**	**96.92**	**12.94**	**95.0**	**84.86**	**57.15**

Table 1: Comparison of various structured pruning approaches at 20% and 50% pruning ratios. All tasks follow zero-shot testing protocol. Best in bold. For fair comparison, we apply recovery training to all methods and report the better results. Full Model-g (3.9B) and Full Model-L (3.2B) for the Vit-g $\mathrm{FlanT5_{XL}}$ and the Vit-L $\mathrm{FlanT5_{XL}}$ version of BLIP-2, respectively.

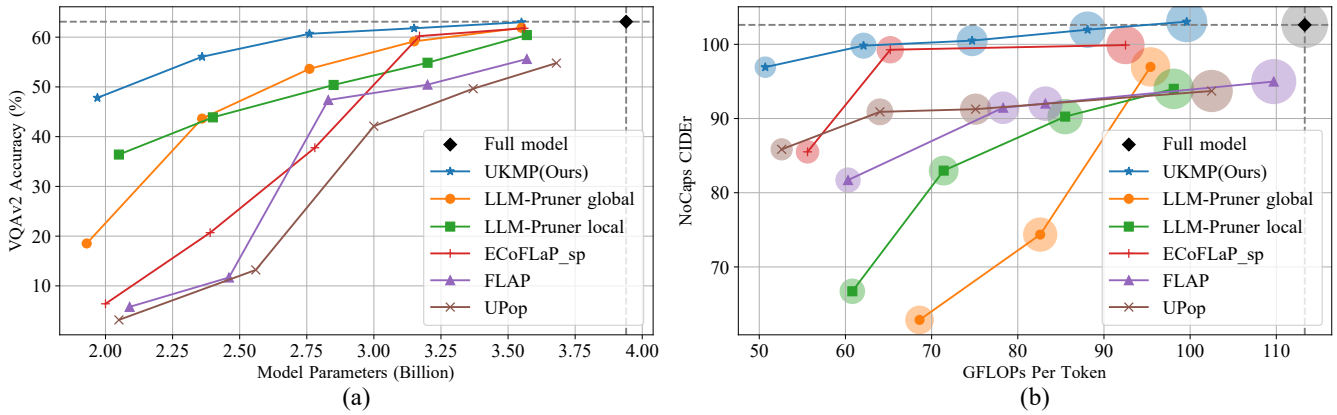

Figure 3: (a) Comparison of various approaches on VQAv2 at different parameter volumes. (b) Comparison of various approaches on NoCaps at different GFLOPs, with the area of circles representing parameter volume. Since the termination times vary across models in auto-regressive generation with beam search, we calculate GFLOPs per token with the same samples.

AntGMM (Wang et al. 2024) is designed for LVLMs without releasing the code. We therefore omit it in comparison.

As summarized in Tab. 1 and Fig. 3, UKMP significantly outperforms the compared methods, with its advantages becoming more pronounced at higher pruning ratios. Concretely, on image captioning, UKMP improves the second-best method by 2.07 and 6.02 in CIDEr at 20% and 50% pruning ratio, respectively. The reason lies in that UKMP addresses the imbalance in unified pruning and adopts a more accurate gradient-based importance criterion. As for VQA, UKMP surpasses the second-best method by 11.41% and 10% at a 50% pruning ratio on VQAv2 and GQA, respectively. This is attributed to UKMP's ability to perceive the correlation between parameters and knowledge capacity, as well as the knowledge recycling from the pruned structures.

Further Remark The baseline LLM-Pruner exhibits severe imbalance with its global scheme and suboptimal performance with its local scheme, suggesting that a globally adaptive pruning is more promising. Following (An et al. 2024), we extend ECoFLaP, an unstructured pruning method for LVLMs, into a structured pruning scheme. It outperforms the other methods except fof UKMP at a 20% pruning ratio, highlighting the importance of addressing modality differences for LVLMs. FLAP claims to have observed structured sample stability in LLMs; however, in LVLMs, the presence of visual encoders renders such properties inapplicable to the entire model. Despite exploring different reconstruction and fine-tuning strategies, the results remain uncompetitive. UPop aims to achieve modality-balanced sub-model search for VLMs, but the comprehensive updating of all parameters disrupts the original knowledge structure, resulting in a substantial performance drop in VQA tasks. Notably, at a 20%

UKMI	LPD	VQAv2	OK-VQA	GQA	NoCaps	
		Accuracy (%)			CIDEr	SPICE
		18.53	9.77	12.01	62.86	10.57
✓		39.01	24.61	30.08	93.39	12.81
✓	✓	**47.81**	**27.38**	**35.15**	**96.92**	**12.94**

Table 2: Effect of the proposed UKMI and LPD modules on VQA and image captioning tasks at 50% pruning ratio.

UN	TS	EI	VQAv2	OK-VQA	GQA	NoCaps	
			Accuracy (%)			CIDEr	SPICE
			29.22	11.73	22.63	68.80	10.89
✓			33.43	23.26	24.58	94.69	12.78
✓	✓		39.54	23.95	29.62	**98.01**	**13.12**
✓	✓	✓	**47.81**	**27.38**	**35.15**	96.92	12.94

Table 3: Ablation study on UKMI at 50% pruning ratio w/ LPD. UN for the unifying on modules and modalities, TS for the Taylor selection over next token prediction, EI for the entropy injection for LLM component.

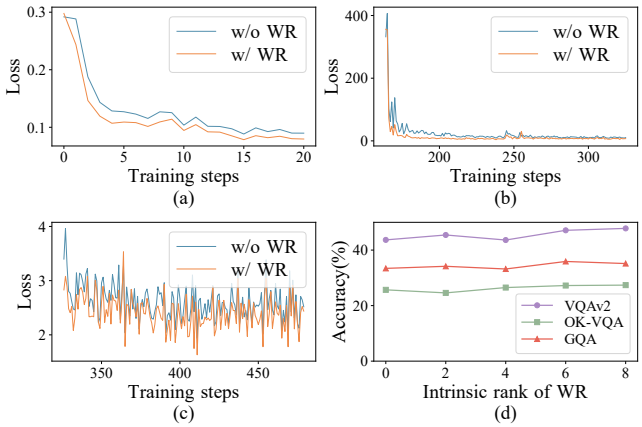

Figure 4: Loss curves w/ and w/o the weight recalling module at 50% pruning ratio in (a) phase τ_1, (b) phase τ_2 and (c) phase τ_3. (d) Ablation study on the intrinsic rank (*i.e.* r_2) of the weight recalling module on VQA tasks.

pruning ratio, our pruned model outperforms the official full model of similar size (*i.e.* BLIP-2 3.2B), even surpassing the original full model on certain tasks (*e.g.* on OK-VQA and GQA), indicating inherent redundancy within LVLMs.

Ablation Study

On Main Components We evaluate the effect of the main modules, *i.e.* UKMI and LPD. As shown in Tab. 2, at a 50% pruning ratio, UKMI prevents the pruned model from nearly collapsing, improving CIDEr on NoCaps by 30.53, while LPD further improves it by 3.53. Both modules promotes the performance in the knowledge-intensive VQA tasks, boosting the VQAv2 accuracy by 20.48% and 8.8%, respectively.

On UKMI We evaluate the effect of the three sub-modules within UKMI. As shown in Tab. 3, the unifying on modules and modalities significantly enhances the overall performance. The Taylor selection further improves image captioning, since it is consistent with the refined gradient-based criterion. The entropy injection boosts the accuracy in VQA, despite a slight decline on image captioning, which is acceptable due to its low reliance on knowledge capacity.

PL	WR	VQAv2	OK-VQA	GQA	NoCaps	
		Accuracy (%)			CIDEr	SPICE
		39.01	24.61	30.08	93.39	12.81
✓		43.67	25.68	33.41	95.64	12.93
✓	✓	**47.81**	**27.38**	**35.15**	**96.92**	**12.94**

Table 4: Ablation study on LPD at 50% pruning ratio w/ UKMI. PL for the progressive learning strategy, WR for the LoRA-based recovery structure with weight recalling.

r_1	r_2	Trainable	VQAv2	OK-VQA	NoCaps	
		Params	Accuracy (%)		CIDEr	SPICE
8	0	18M	43.67	25.68	95.64	12.93
4	4	17M	45.81	27.30	96.24	12.88
16	0	35M	43.57	25.38	95.83	**12.96**
8	8	33M	**47.81**	**27.38**	**96.92**	12.94

Table 5: Ablation study on trainable parameters in LPD.

On LPD We evaluate the effect of the two sub-modules within LPD. As shown in Tab. 4, both the LoRA-based weight recalling structure and the progressive learning strategy promote the performance. Loss curves in Fig. 4 indicates that the weight recalling module improves convergence during training. Despite doubling the number of trainable parameters, the ablation results in Tab. 5 reveal that adding more parameters to the original LoRA module yields slight improvements. Instead, the significant gains are primarily attributed to the proposed weight recalling module. Additionally, Fig. 4(d) shows that a higher intrinsic rank of weight recalling module (*e.g.* 6, 8) facilitates better knowledge recovery, whereas excessively small ranks (*e.g.* 2, 4) can have detrimental effects.

Conclusion

In this paper, we investigate unified structured pruning for LVLMs, and propose a novel approach dubbed Unified Knowledge Maintenance Pruning and Progressive Recovery with Weight Recalling (UKMP). We develop a unified knowledge maintenance importance metric to refine the gradient-based criterion, perceive the model's knowledge capacity and resolve imbalance for globally unified pruning. Moreover, a LoRA-based progressive distillation is employed to facilitate recovery with limited data. Experiments on various vision-language tasks demonstrate the effectiveness of our approach. Notably, the latest LVLMs increasingly rely on more complex instruction data to build ultra-large models (Liu et al. 2023). Handling the redundancy within these models remains a challenging problem. We hope our contributions will assist in advancing the understanding and exploration of pruning techniques for LVLMs.

Acknowledgements

This work was partly supported by the Beijing Municipal Science and Technology Project (No. Z231100010323002), the National Natural Science Foundation of China (No. 62202034), the Beijing Natural Science Foundation (No. 4242044), the Aeronautical Science Foundation of China (2023Z071051002), the Research Program of State Key Laboratory of Virtual Reality Technology and Systems, and the Fundamental Research Funds for the Central Universities.

References

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8947–8956.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. In *Proceedings of the Neural Information Processing Systems*, 23716–23736.

An, Y.; Zhao, X.; Yu, T.; Tang, M.; and Wang, J. 2024. Fluctuation-based adaptive structured pruning for large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 10, 10865–10873.

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision*, 382–398.

Ashkboos, S.; Croci, M. L.; Nascimento, M. G. d.; Hoefler, T.; and Hensman, J. 2024. Slicegpt: Compress large language models by deleting rows and columns. In *Proceedings of the International Conference on Learning Representations*.

Cao, J.; Ye, P.; Li, S.; Yu, C.; Tang, Y.; Lu, J.; and Chen, T. 2024. Madtp: Multimodal alignment-guided dynamic token pruning for accelerating vision-language transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15710–15719.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70): 1–53.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. C. H. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Proceedings of the Neural Information Processing Systems*.

Fang, C.; Zhou, A.; and Wang, Z. 2022. An algorithm–hardware co-optimized framework for accelerating n: m sparse transformers. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 30(11): 1573–1586.

Fang, G.; Ma, X.; Song, M.; Mi, M. B.; and Wang, X. 2023. Depgraph: Towards any structural pruning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16091–16101.

Fang, G.; Ma, X.; and Wang, X. 2023. Structural pruning for diffusion models. In *Proceedings of the Neural Information Processing Systems*.

Frantar, E.; and Alistarh, D. 2023. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *Proceedings of the International Conference on Machine Learning*, 10323–10337.

Frantar, E.; Ashkboos, S.; Hoefler, T.; and Alistarh, D. 2023. Gptq: Accurate post-training quantization for generative pre-trained transformers. In *Proceedings of the International Conference on Learning Representations*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. 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*, 6904–6913.

He, S.; Li, A.; and Chen, T. 2024. Rethinking Pruning for Vision-Language Models: Strategies for Effective Sparsity and Performance Restoration. arXiv:2404.02424.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*.

Hudson, D. A.; and Manning, C. D. 2019. 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*, 6700–6709.

Jiang, L.; Chen, J.; Huang, D.; and Wang, Y. 2023. MIEP: Channel Pruning with Multi-granular Importance Estimation for Object Detection. In *Proceedings of ACM International Conference on Multimedia*, 2908–2917.

Jin, Y.; Li, J.; Liu, Y.; Gu, T.; Wu, K.; Jiang, Z.; He, M.; Zhao, B.; Tan, X.; Gan, Z.; Wang, Y.; Wang, C.; and Ma, L. 2024. Efficient multimodal large language models: A survey. arXiv:2405.10739.

Kim, B.-K.; Kim, G.; Kim, T.-H.; Castells, T.; Choi, S.; Shin, J.; and Song, H.-K. 2024. Shortened LLaMA: Depth Pruning for Large Language Models with Comparison of Retraining Methods. arXiv:2402.02834.

LeCun, Y.; Denker, J.; and Solla, S. 1989. Optimal brain damage. In *Proceedings of the Neural Information Processing Systems*, 598–605.

Li, D.; Li, J.; Le, H.; Wang, G.; Savarese, S.; and Hoi, S. C. 2023a. Lavis: A library for language-vision intelligence. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 31–41.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*, 19730–19742.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning*, 12888–12900.

Liang, T.; Glossner, J.; Wang, L.; Shi, S.; and Zhang, X. 2021. Pruning and quantization for deep neural network acceleration: A survey. *Neurocomputing*, 461: 370–403.

Lin, H.; Bai, H.; Liu, Z.; Hou, L.; Sun, M.; Song, L.; Wei, Y.; and Sun, Z. 2024. Mope-clip: Structured pruning for efficient vision-language models with module-wise pruning error metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27370–27380.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. In *Proceedings of the Neural Information Processing Systems*.

Liu, H.; Xue, W.; Chen, Y.; Chen, D.; Zhao, X.; Wang, K.; Hou, L.; Li, R.; and Peng, W. 2024. A survey on hallucination in large vision-language models. arXiv:2402.00253.

Ma, X.; Fang, G.; and Wang, X. 2023. Llm-pruner: On the structural pruning of large language models. In *Proceedings of the Neural Information Processing Systems*, 21702–21720.

Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3195–3204.

Pi, R.; Gao, J.; Diao, S.; Pan, R.; Dong, H.; Zhang, J.; Yao, L.; Han, J.; Xu, H.; Kong, L.; and Zhang, T. 2023. Detgpt: Detect what you need via reasoning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 14172–14189.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, 2641–2649.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2556–2565.

Shi, D.; Tao, C.; Jin, Y.; Yang, Z.; Yuan, C.; and Wang, J. 2023. Upop: Unified and progressive pruning for compressing vision-language transformers. In *Proceedings of the International Conference on Machine Learning*, 31292–31311.

Sun, M.; Liu, Z.; Bair, A.; and Kolter, J. Z. 2023a. A simple and effective pruning approach for large language models. In *Proceedings of the International Conference on Learning Representations*.

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

Sung, Y.-L.; Yoon, J.; and Bansal, M. 2024. Ecoflap: Efficient coarse-to-fine layer-wise pruning for vision-language models. In *Proceedings of the International Conference on Learning Representations*.

Team, G.; Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; et al. 2024. Gemini: a family of highly capable multimodal models. arXiv:2312.11805.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4566–4575.

Wang, M.; Zhao, Y.; Liu, J.; Chen, J.; Zhuang, C.; Gu, J.; Guo, R.; and Zhao, X. 2024. Large multimodal model compression via iterative efficient pruning and distillation. In *Companion Proceedings of the ACM on Web Conference*, 235–244.

Wu, Z.; Chen, J.; and Wang, Y. 2023. SAMP: Sub-task Aware Model Pruning with Layer-Wise Channel Balancing for Person Search. In *Proceedings of the Chinese Conference on Pattern Recognition and Computer Vision*, 199–211.

Wu, Z.; Chen, J.; Zhong, H.; Huang, D.; and Wang, Y. 2025. AdaLog: Post-training Quantization for Vision Transformers with Adaptive Logarithm Quantizer. In *Proceedings of the European Conference on Computer Vision*, 411–427.

Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; and Chen, E. 2024. A survey on multimodal large language models. arXiv:2306.13549.

Zhang, Q.; Zuo, S.; Liang, C.; Bukharin, A.; He, P.; Chen, W.; and Zhao, T. 2022. Platon: Pruning large transformer models with upper confidence bound of weight importance. In *Proceedings of the International Conference on Machine Learning*, 26809–26823.