



METEOR: Multi-Encoder Collaborative Token Pruning for Efficient Vision Language Models

Yuchen Liu¹, Yaoming Wang^{3†}, Bowen Shi¹, Xiaopeng Zhang^{2†}, Wenrui Dai^{1†}, Chenglin Li¹,
Hongkai Xiong¹ and Qi Tian²

¹Shanghai Jiao Tong University, China ²Huawei Inc., China ³Meituan Inc., China

{liuyuchen6666, sjtu.shibowen, daiwenrui, lcl1985, xionghongkai}@sjtu.edu.cn

wangyaoming03@meituan.com, zxphistory@gmail.com, tian.qil@huawei.com

Abstract

Vision encoders serve as the cornerstone of multimodal understanding. Single-encoder architectures like CLIP exhibit inherent constraints in generalizing across diverse multimodal tasks, while recent multi-encoder fusion methods introduce prohibitive computational overhead to achieve superior performance using complementary visual representations from multiple vision encoders. To address this, we propose a progressive pruning framework, namely **Multi-Encoder Collaborative Token Pruning (METEOR)**, that eliminates redundant visual tokens across the encoding, fusion, and decoding stages for multi-encoder MLLMs. For multi-vision encoding, we discard redundant tokens within each encoder via a rank guided collaborative token assignment strategy. Subsequently, for multi-vision fusion, we combine the visual features from different encoders while reducing cross-encoder redundancy with cooperative pruning. Finally, we propose an adaptive token pruning method in the LLM decoding stage to further discard irrelevant tokens based on the text prompts with dynamically adjusting pruning ratios for specific task demands. To our best knowledge, this is the first successful attempt that achieves an efficient multi-encoder based vision language model with multi-stage pruning strategies. Extensive experiments on 11 benchmarks demonstrate the effectiveness of our proposed approach. Compared with EAGLE, a typical multi-encoder MLLMs, **METEOR** reduces 76% visual tokens with only 0.3% performance drop in average. The code is available at <https://github.com/YuchenLiu98/METEOR>.

1. Introduction

Recent advancements in large language models (LLMs) [2, 4, 12, 63] have significantly propelled the development of

[†]Corresponding author: Yaoming Wang, Xiaopeng Zhang, Wenrui Dai. This work was done when Yuchen Liu interned at Huawei Inc.

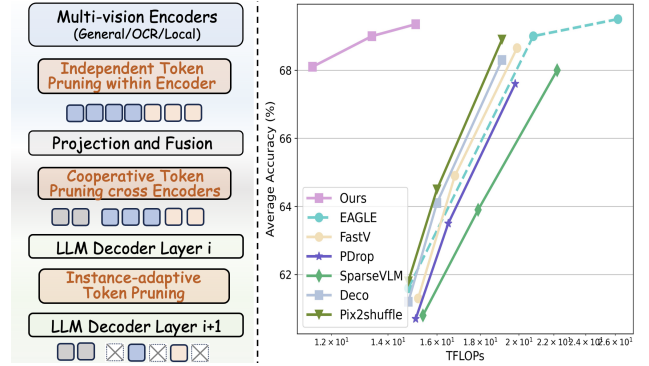


Figure 1. Overview of METEOR, which collaboratively prune redundant visual tokens across the encoding, fusion and decoding process of Multi-encoder MLLMs. Comparison with existing efficient methods based on EAGLE-X4 [58] regarding accuracy and flops. Other two results of EAGLE are with one and two encoders.

multi-modal LLMs (MLLMs), enabling promising applications that bring artificial intelligence into daily life [3, 9, 35, 40, 82]. Typical MLLMs transform images into sequences of visual tokens, and then concatenate them with text tokens for LLMs. Subsequent works further enhance performance through high-resolution inputs [16, 36] and innovative transformer architectures [3, 61]. However, due to the inherently sparse nature of visual information compared to textual information, directly concatenating visual tokens causes significant computational inefficiency. To mitigate this inefficiency, prompt-agnostic methods reduce the number of visual tokens before feeding into LLMs for single-encoder MLLMs like LLaVA [40]. Visual tokens are compressed by visual attention mechanisms [25, 57, 75], local pooling [5, 13, 35, 69], and resamplers [16]. However, these methods overlook the context provided by text prompts, resulting in redundant visual tokens irrelevant to user instructions. As a remedy, prompt-aware methods [20, 67, 71, 74] such as FastV [7], Pdrop [67] and SparseVLM [78] employ text prompts to guide the visual token pruning in LLMs.

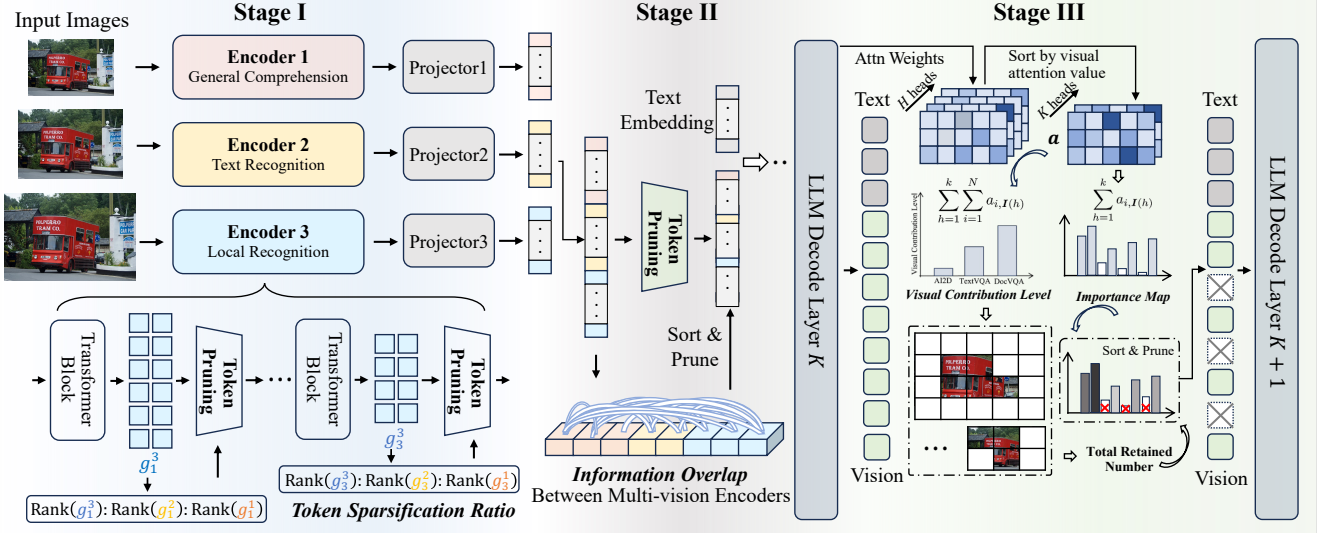


Figure 2. Overall architecture of METEOR. Redundant visual tokens are pruned within each encoder during encoding in stage 1. The rank of output feature map is utilized to assess the information richness for allocating the token sparsity ratio of each encoder. Besides, based on the independent projector for adapting tokens to the shared semantic space, we prune mutually redundant visual tokens with information overlap across multi-encoders. Finally, we employ the vision-text attention value of the top- k most relevant attn heads to precisely identify redundant tokens, and dynamically adjust the retained token number for varying task demands based on visual contribution level.

On the other hand, widely adopted single-encoder architectures like CLIP [56] suffer from hallucinations [36, 42] and are degraded on fine-grained tasks such as grounding and OCR [24, 66]. To address these limitations, recent works [14, 27, 43, 58, 61, 84] explore integrating multiple state-of-the-art vision encoders to improve visual perception for robust performance across diverse domains. Despite promising performance, multiple vision encoders with high-resolution inputs introduce prohibitive computational overhead. For instance, in Mini-Gemini [35], a 672×672 image processed by dual vision encoders generates 2880 visual tokens, leading to quadratic scaling of computational complexity in self-attention mechanisms.

To tackle this problem, one intuitive approach is to apply single-encoder token pruning strategies in multi-encoder MLLMs. However, such an approach falls short in achieving collaborative pruning across multiple vision encoders. Specifically, it fails to address critical issues such as the appropriate allocation of token pruning ratios to vision encoders with varying levels of semantic richness and effective reduction of redundant tokens that overlap across multiple vision encoders. It is also crucial to accelerate the whole process under the complexity of vision encoding increasing with image resolution in Figure 3(a). Moreover, existing pruning methods perform poorly on some specific tasks like OCR recognition due to their inability to dynamically adjust pruning ratio for varying task demands, since OCR tasks require retaining more tokens than general tasks.

In this paper, we propose a novel multi-modal token pruning framework, namely **Multi-Encoder Collaborative**

token pRuning (METEOR) to progressively eliminate redundant visual tokens across the whole process of multi-encoder MLLMs, including encoding, fusion and decoding. Specifically, for multi-vision encoding, we identify reliable criterion that measures similarity to the average token in shallow layers and class attention in deep layers for progressive pruning of redundant visual tokens. To allocate the appropriate token sparsity ratio for each encoder, we leverage the rank of feature maps produced by different encoders as the mathematically grounded measure of information richness. Furthermore, we find that the average rank of feature maps in a layer of the encoder is robust to input images with small variance, and provide tractable offline rank computation on a small batch with negligible computation overhead.

For multi-vision encoder fusion, we employ a more flexible fusion strategy to maintain its own dedicated projector for each vision encoder, enabling independent adapting of visual tokens before fusion. Based on the adapted visual embeddings, we propose a cooperative pruning strategy to reduce mutually redundant visual tokens with information overlap across multi-vision encoders. Finally, for LLM decoding, we employ visual-text attention to reduce redundant visual tokens irrelevant to the text prompt. Different from existing methods averaging all the attention heads, we select top- k significant heads to mitigate hallucination. Leveraging the correlation between the visual attention value and instance complexity across datasets, we dynamically adjust the token pruning ratio for specific tasks like OCR.

To our best knowledge, METEOR is the first token pruning framework tailored for multi-encoder MLLMs. The

contributions are summarized as:

- We propose a multi-encoder collaborative token pruning framework to progressively eliminate redundant visual tokens across the encoding, fusion, and decoding stages.
- We propose to utilize the rank of feature maps as a mathematically grounded measure of information richness to collaboratively allocate sparsity ratios for each encoder, and prune mutually redundant tokens with information overlap across multi-encoders during fusion.
- We develop an instance-adaptive pruning strategy that selects most relevant attention heads for accurate visual-text attention to identify redundant tokens, and further dynamically adjusts pruning ratios for varying task demands.

Extensive experiments on multimodal benchmarks demonstrate the effectiveness of METEOR. Compared with state-of-the-art multi-encoder MLLMs EAGLE, METEOR reduces 76% visual tokens while saving 49% TFLOPS and accelerating 46% FPS with only 0.3% performance drop. Besides, our method consistently outperforms prior token pruning methods by 4.3-5.7% on average, especially 8.8-12.3% on OCR tasks, exhibiting the advantage of our instance-adaptive token pruning. Moreover, with the same vision encoders, our method significantly outperforms Cambrian-1 [61] by 3.3% on OCR tasks with 44% fewer visual tokens, demonstrating the broad applicability of our strategy across various multi-encoder MLLMs.

2. Related Work

Multi-modal Large Language Models (MLLMs). In recent years, MLLMs [3, 9, 35, 40, 61] have achieved promising progress. LLaVA [40] and MiniGPT-4 [82] conduct instruction tuning on high-quality datasets to enhance generation capabilities under complex instructions. For further improvement, later researches explore high-resolution inputs [16, 36, 42], model design [3, 61], and scaling up model size and data [8, 9, 35]. However, processing high-resolution images inevitably results in an exponential increase in the length of visual tokens. Since the number of text tokens is much smaller, the overall computational cost grows quadratically with the number of visual tokens. Thus, how to reduce the redundant visual tokens while maintaining the performance is critical for improving efficiency.

Vision Encoders Design for MLLMs. Previous MLLMs usually adopt typical vision-language pretrained models CLIP [56, 60]. Despite good abilities in general recognition, these models lack specific abilities such as reading text and localizing objects. Hence, a series of works [18, 24, 29, 35, 38, 43, 48, 62] have integrated vision models pre-trained on diverse vision tasks or vision-language tasks to extend the visual capabilities. For example, Mousi [14], Brave [27], Cambrian-1 [61] and EAGLE [58] fuse visual tokens from different vision experts, e.g., OCR, detection and segmentation, by concatenating along the channel or token direction.

MoVA [84] proposes a routing network to select optimal vision model combinations based on the given instructions. Despite promising performance, the visual encoding time of multi-vision encoders increases dramatically, even several times LLM prefilling time. Thus, how to reduce the computation complexity of multi-vision encoders in MLLMs with an optimized accuracy-latency trade-off is an urgent issue.

Visual Token Compression for MLLMs. Compared with text, visual tokens occupy the majority of the input sequence with high redundancy, making visual token compression effective for accelerating MLLMs. Prompt-agonistic methods [1, 22, 25, 57, 64, 65, 75] leverage attention mechanisms of the vision encoder to identify important visual tokens. DeCo [69] and TokenPacker [33] reduce visual tokens with efficient visual projectors as pooling or local attention. Prompt-aware methods [19, 71, 72, 74, 80, 81, 83] leverage the text semantics to guide visual token pruning in LLMs. FastV [7] prunes unnecessary visual tokens based on the vision-text attention values. PDrop [67] and Sparse-VLM [78] progressively reduce the number of retained visual tokens as the LLM layers deepen. The visual computation of self-attention and FFN is saved in [32, 79]. LLaVA-Mini [76] and VoCo-LLaMA [73] explore extreme compression to one visual token. MQT [21] and M3 [5] employ matryoshka representation learning to compress visual tokens. iLLaVA [20], Ficoco [17] and MustDrop [44] propose to drop visual tokens both in vision encoding stage and LLM stage. Despite promising results, these methods cannot be directly applied to multi-encoder MLLMs, e.g., how to allocate token budgets to various vision encoders. Besides, existing methods fail to adaptively adjust token budget for varying tasks, which perform poorly on OCR tasks.

3. Methodology

3.1. Overview

Figure 2 depicts the proposed framework that progressively prunes redundant visual tokens in the whole process of multi-vision encoding, multi-vision fusion, and LLM decoding stages for multi-encoder MLLMs.

Stage 1: Multi-vision Encoding. Figure 3(a) shows that the latency for vision encoding dramatically increases with the growth of image resolution and model complexity, and could even exacerbate for multi-vision encoders. We develop a reliable measure to identify and prune redundant tokens with a collaboratively allocated sparsity ratio.

Stage 2: Multi-vision Fusion. Existing methods [58, 61] adapt the independently pruned tokens from different encoders with one shared projector. They are not flexible and overlook mutually redundant tokens with information overlap across multi-encoders. We employ the dedicated projector to adapt tokens from each encoder and cooperatively suppress mutual redundancy across encoders during fusion.

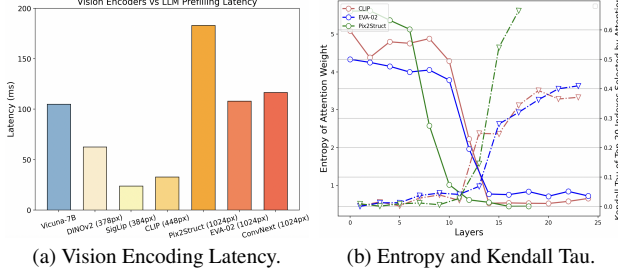


Figure 3. (a) Vision encoding and LLM prefilling latency (fed with 576 tokens). (b) Entropy of attention value and kendall tau correlation of top- k indices selected by attention values between adjacent layers.

Stage 3: LLM Decoding. Pruning using only vision information cannot remove redundancy related to specific text prompts, nor fulfill different token budget for instances of varying complexity, *e.g.*, more tokens required for OCR than general tasks. We adaptively prune visual tokens with dynamically adjusted pruning ratio guided by text prompts.

3.2. Stage 1: Independent Pruning within Encoders

Given L vision encoders with the l -th encoder stacking B_l blocks, we propose to progressively discard tokens in multiple blocks to sufficiently exploit the information propagation capability of self-attention mechanism and significantly alleviate the serious information loss by pruning all at once. For any b -th block in the l -th vision encoder, we figure out the redundant tokens and allocate a proper pruning ratio to each encoder with varying information richness.

Redundant Token Identification. Attention values are usually used to calculate the significance of each token [25, 57, 64, 75] but are not reliable in shallow layers. As shown in Figure 3(a), the distribution of attention values is not sparse in shallow layers and exhibits high entropy. Besides, the top- k indices selected by attention values vary significantly with low Kendall’s tau correlation between adjacent layers. Since shallow layers usually contain low-level information, the average token corresponding to low-frequency components using discrete Fourier transform usually represents the background with high redundancy. Thus, we measure the significance with similarity to the average token as

$$T(g_b^l; k) = \text{gather}[g_b^l, \text{sort}(-S(g_b^l, \sum_{n=1}^n g_b^l/n)) < k]. \quad (1)$$

Here, $g_b^l \in \mathbb{R}^{N_b^l \times D}$ is the output tokens from the b -th block in the l -th vision encoder, sort means sorting descendingly, $S(\cdot, \cdot)$ is the cosine similarity, and $T(g_b^l; k)$ is the set of retained k tokens.

Finding 1: Average token is more suitable than attention values for measuring the token significance in shallow blocks containing low-level redundant information.

Attention values are sparse and reliable for deep layers, as shown in Figure 3(b). We measure the significance with

attention values between the class token and visual tokens.

$$T(g_b^l; k) = \text{gather}[g_b^l, \text{sort}(\text{softmax}(\mathbf{q}_{\text{cls}} \cdot \mathbf{K}^T / \sqrt{d_k}) < k)], \quad (2)$$

where $\mathbf{q}_{\text{cls}}$ is the query value of the cls token and $\mathbf{K}$ is the key value of g_b^l . In practice, we equally split the model into three phases, where the first shallow phase employs the cosine similarity of features and last two phases employ the reliable attention value to identify redundant visual tokens.

Sparsity Ratio for Multi-vision Encoders. Visual tokens from different vision encoders exhibit varying levels of semantic richness and redundancy, and differ in contributions to MLLM understanding. Here, we allocate varying sparsity ratios for each block to appropriately budget the number of retained tokens for semantic richness, *i.e.*, fewer tokens for lower semantic richness. Inspired by [37], we explore the rank of the feature map as a measure of token redundancy. The singular value decomposition (SVD) of g_b^l is

$$g_b^l = \sum_{b=1}^r \sigma_b \alpha_b w_b^\top = \sum_{b=1}^{r'} \sigma_b \alpha_b w_b^\top + \sum_{b=r'+1}^r \sigma_b \alpha_b w_b^\top. \quad (3)$$

where r is the rank of g_b^l , and α_b, w_b are the top- i left and right singular vectors. A feature map of rank r is decomposed into a feature map $\sum_{b=1}^{r'} \sigma_b \alpha_b w_b^\top$ with lower rank r' and additional information $\sum_{b=r'+1}^r \sigma_b \alpha_b w_b^\top$. This implies that rank is a reliable measure of information richness, since feature maps with higher rank contain more information.

Since online computation of ranks is prohibitive, we use a small batch of input images to estimate the expectation of ranks in an offline manner. We extend the findings in [37] from CNNs to Vision Transformers and empirically reveal that the expectation of ranks is robust to the input images and the variance is negligible, as shown in Figure 4(a).

Finding 2: The rank of feature map is a stable and reliable measure for the information richness of various encoders.

The retained token number budget for the b -th layer of the l -th encoder is allocated proportionally to the rank r_b^l as $k_b^l = k_b \cdot r_b^l / \sum_{l=1}^C r_b^l$, where k_b is the overall kept token number of the b -th block for all vision encoders.

3.3. Stage 2: Cooperative Pruning cross Encoders

Multi-vision Token Fusion. Existing multi-encoder MLLMs [27, 50, 58] employ a pre-projection fusion strategy to combine visual tokens before the vision-text alignment using a shared projector for all the vision encoders. In contrast, we adopt a more flexible post-projection fusion where each encoder maintains its own dedicated projector to independently adapt tokens for fusion. This strategy could align visual embeddings from various encoders to further reduce redundant tokens across multi-encoders in the shared semantic space. Specifically, we use a two-layer MLP G as the projector, *i.e.*, $z^j = G(g^l)$, where g^l is the output of the

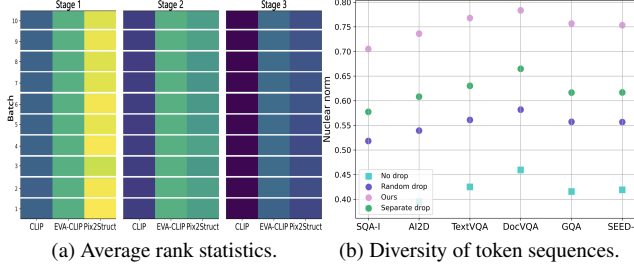


Figure 4. (a) Rank statistics of feature maps across different phases of encoders, with color indicating rank values. As shown, the rank of each feature map hardly changes regardless of image batches. (b) Diversity of concatenated token sequences in stage 2.

l -th encoder. Visual tokens from multi-encoders are concatenated along the token channel as $z = [z^1, \dots, z^L]$.

Mutual Redundancy across Multi-vision Encoders. Tokens with similar representations tend to have analogous contributions in the following attention mechanism and lead to mutual redundancy. We further achieve cooperative pruning to remove mutual redundancy across various vision encoders for better efficiency. For the i -th token z_i^j in z^j , the mutual redundancy is defined as:

$$\mathcal{R}_i^j = \sum_{l=1, l \neq j}^L \sum_{m=1}^{n_l} \mathcal{S}(z_i^j, z_m^l). \quad (4)$$

We then keep the top- k tokens with the lowest mutual redundancy and discard the rest as $T(z; k) = \text{gather}[z, \text{sort}(-\mathcal{R}) < k]$.

Analysis on Diversity Enhancement. We measure the feature diversity with the nuclear norm (more details in the supplementary material). Figure 4(b) shows our cooperative pruning effectively enhances the diversity of token sequences compared to no or random discarding. Moreover, it outperforms separately dropping tokens in each encoder, highlighting the importance of reducing cross-encoder redundancy over within-encoder redundancy during fusion.

Finding 3: Reducing mutual redundancy across feature tokens of multi-vision encoders achieves better efficiency.

3.4. Stage 3: Text-aware Instance-adaptive Pruning

Redundant Token Identification. Redundant tokens can be further reduced considering specific text prompts beyond visual information and the sparsity ratio can be dynamically adjusted for varying complexity of input instances. Existing methods [7, 20, 67, 71] suffer from degraded performance on fine-grained tasks using a pre-defined sparsity ratio. We identify redundant visual tokens with the attention value a_i between the visual token z_i and the last instruction token t .

Finding 4: Not all attention heads are relevant for accurately identifying redundant visual tokens.

We adopt the Visual Attention Value (VAV) that is the magnitude of the attention value between the text token t

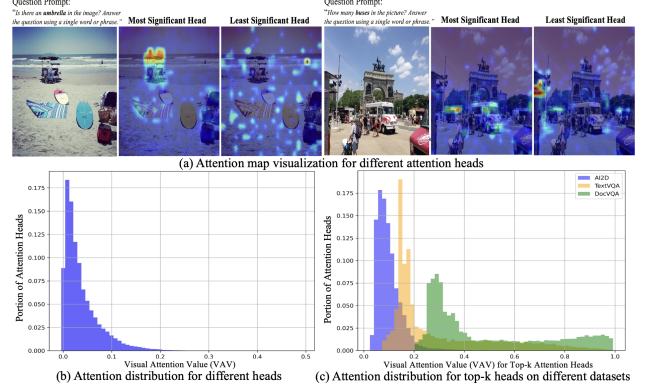


Figure 5. (a) Attention map visualization for different heads of the most significant value and the least value. (b) Visual attention value for different attention heads on AI2D. (c) Visual attention value for top- k largest attention heads on different datasets.

and all visual tokens to measure the quality of each attention head. For the h -th head,

$$\text{VAV}_h \triangleq \sum_{i=1}^N a_{i,h} \quad (5)$$

where $a_{i,h}$ is the attention value of the h -th attention head. Existing methods [7, 20, 67, 72, 78] compute the attention value by averaging all the attention heads, and could focus on the wrong visual area with hallucination [26, 68, 77], as shown in Figure 5(a). On the contrary, we find that not all the self-attention heads in LLMs are relevant to accurately grounding the text tokens to visual tokens. Figure 5(b) shows that VAV_h varies a lot across attention heads, and most heads are close to zero, which are irrelevant. Since heads with higher value are more reliable, we sort and retain the top- k most significant heads as $\mathbf{I} = \arg \max(\text{VAV}_1, \dots, \text{VAV}_H)[k]$, and derive the importance criteria as $a'_i = \sum_{h=1}^k a_{i,\mathbf{I}(h)}$ to identify redundant visual tokens. Then the visual tokens are sorted and tokens with large importance values are retained while the rest are pruned out in successive layers.

Instance-adaptive Token Retention. We find the VAV of top- k most significant attention heads correlates with the complexity of varying inputs, and thus propose to dynamically adjust the retained token numbers based on VAV.

Finding 5: The distribution of Visual Attention Value varies a lot for different datasets, and is closely related to the complexity of input instances.

Figure 5(c) exhibits a clear correlation between the VAV and instance complexity across datasets. Specifically, AI2D characterized by coarse-grained general comprehension exhibits low value, while DocVQA demanding fine-grained OCR recognition shows high VAV. Actually, the VAV quantifies the degree to which text tokens interact with visual information, with higher values signifying a greater contribution of image tokens during generation. Thus, we define the

Method	LLM	Eff. Res.	Vis. Tok.	PT	SFT	Knowledge			General				OCR and Chart			
						SQA	AI2D	OK VQA	GQA	POPE	SEED	MMB	Text VQA	Doc VQA	Chart QA	OCR Bench
InstructBLIP [11]	Vic.7B	224	32	129M	1.2M	60.5	-	-	49.2	-	53.4	-	50.1	-	-	291
Qwen-VL [3]	Qwen	448	256	1.4B	50M	67.1	-	-	59.3	-	56.3	38.2	61.5	65.1	-	-
Qwen-VL-Chat [3]	Qwen	448	256	1.4B	50M	68.2	-	-	57.5	-	58.2	60.6	63.8	62.6	-	488
mPLUG-Owl2 [70]	LLa-2	448	1024	400M	1.2M	68.7	-	-	56.1	-	57.8	64.5	54.3	-	-	366
MM1 [54]	-	1344	720	3B	1.5M	62.3	-	-	-	87.4	65.6	-	68.2	68.4	-	-
LLaVA-Next [42]	Vic.7B	672	2880	558K	765K	70.1	66.6	44.3	64.2	86.5	70.2	67.4	64.9	70.0	64.0	532
Mini-Gemini [35]	Vic.7B	1536	2880	1.5M	1.5M	-	-	-	-	86.8	-	65.8	68.4	65.0	-	456
SPHINX-2k [38]	Vic.13B	762	2890	-	-	70.6	-	-	63.1	87.2	-	65.9	61.2	-	-	-
Pyramiddrop [67]	Vic.7B	672	1169	558K	1.0M	-	-	-	63.7	86.5	69.5	68.1	67.7	66.6	63.0	-
LLaVA-HR [50]	LLa-2	1024	1024	558K	1.2M	65.1	-	-	64.2	87.6	64.2	-	67.1	-	-	-
EAGLE [58]	Vic.7B	1024	1024	558K	1.8M	70.8	-	-	64.9	88.7	73.4	68.1	71.9	-	67.9	544
EAGLE† [58]	Vic.7B	1024	1024	558K	1.8M	71.0	72.2	58.8	64.8	88.4	73.5	68.0	71.7	73.2	67.4	538
TokenPacker [33]	Vic.7B	1088	954	1.5M	1.5M	-	-	-	-	88.3	-	67.4	68.0	60.2	-	452
Mousi [14]	Vic.7B	336	576	1.2M	1.6M	69.0	-	-	63.6	86.5	67.5	67.4	58.4	-	-	-
DeepStack-V [55]	Vic.7B	672	576	558K	665K	-	-	-	64.1	87.6	62.3	-	63.5	41.0	21.0	-
Brave-X5 [27]	FlanT5	224	160	100M	-	-	-	-	52.7	87.6	-	-	-	-	-	-
METEOR (Ours)	Vic.7B	1024	242*	558K	1.8M	71.4	73.4	58.6	63.5	87.9	72.8	69.8	71.1	71.4	65.6	533
METEOR (Ours)	LLa-3	1024	242*	558K	1.8M	77.1	76.9	59.0	63.5	88.2	74.0	72.5	70.8	72.0	68.0	542
METEOR (Ours)	Vic.7B	1024	126*	558K	1.8M	70.7	73.9	58.6	62.9	87.6	71.7	69.7	69.4	67.9	64.3	517
METEOR (Ours)	LLa-3	1024	126*	558K	1.8M	78.2	77.1	59.0	64.3	87.8	73.1	72.5	69.0	68.3	66.5	525

Table 1. Comparison with other MLMMs on 11 benchmarks. *Eff. Res.* indicates the effective image resolution. *Vis. Tok.* indicates the number of visual tokens used in LLMs. “Vic.”, “LLa-2” and “LLa-3” refer to Vicuna, LLaMA-2 and LLaMA-3. † denotes that we reproduce the results with the officially released code and training data (w/o pre-alignment stage). Token with (*) means the retained token count is not pre-defined and adaptively determined, which is calculated by averaging across all benchmarks during inference.

visual contribution level as the sum of VAV of top- k heads $\sum_{h=1}^k \sum_{i=1}^N a_{i,I(h)}$. This increased contribution, in turn, necessitates retaining a larger number of visual tokens during generation. Consequently, the visual contribution level determines the number of visual tokens to retain as:

$$K = \lambda \cdot \sum_{h=1}^k \sum_{i=1}^N a_{i,I(h)} \quad (6)$$

where λ is a constant for scaling. Besides, since the redundancy of visual tokens gradually increases as the layer deepens [7], we progressively compress tokens at three phases.

4. Experiments

4.1. Experimental Settings

Implementation Details. We adopt EAGLE [58] as the base settings, where Vicuna-v1.5-7B [10] and Llama3-8B [63] are used as the LLM, and four vision encoders: CLIP [56], ConvNeXt [47], Pix2Struct [30] and EVA-02 [60] are adopted. For the pre-training stage, we adopt the token discarding strategy in multi-vision encoding and use the same pre-training data as LLaVA-1.5 [41] with 558k image-text pairs, where the whole model is frozen and only the projector is updated. For the supervised fine-tuning, based on the pre-trained projector for alignment, we incorporate cooperative pruning in multi-vision fusion with 576 tokens retained and employ the data recipe in [6] with

1M image-text pairs as the base setup to finetune the whole model, and also employ EAGLE-1.8M [58] data recipe as the advanced setup. Finally, the proposed instance-adaptive Text-guided Pruning is conducted in a training-free manner with different token budget configurations for simplicity. More details are in the supplementary material.

Evaluation Datasets. We evaluate the model on existing diverse multimodal benchmarks including SEEDBench [31], POPE [34], TextVQA [59], ChartQA [52], DocVQA [53], GQA [23], ScienceQA [49], AI2D [28], OCRBench [45], OKVQA [51], MME [15] and MMBench [46].

4.2. Main Results

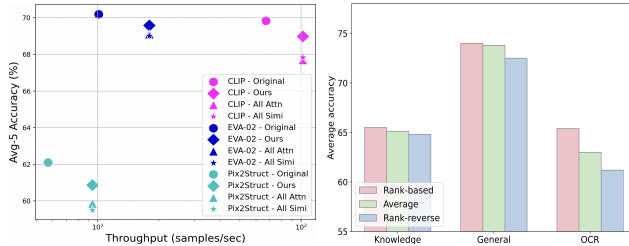
Comparison with Leading MLLMs. As shown in Table 1, we present the performance of our method on 11 vision-language benchmarks with two visual token count configurations, i.e., around 242 and 126 tokens. Compared with existing MLLMs, METEOR consistently achieves superior performance, outperforming LLaVA-Next with less than 10% of the tokens in 10 out of 11 benchmarks. Compared with existing training-based token pruning approaches such as Pyramiddrop, TokenPacker and DeepStack-V, our method demonstrates superior performance with 60%-80% fewer tokens, particularly in more challenging fine-grained OCR recognition, e.g., outperforms by 2.4%, 8.8% and 28% on DocVQA, respectively. Compared with existing multi-expert MLLMs, METEOR significantly outperforms Mousi

Method	Vis. Tok.	TFOLPS	Throughput	Knowledge			General				OCR and Chart				Avg.
				SQA	AI2D	OK VQA	GQA	POPE	SEED	MMB	Text VQA	Doc VQA	Chart QA	OCR Bench	
EAGLE [†] [58]	1024	26.21	0.81	71.0	72.2	58.8	64.8	88.4	73.5	68.0	71.7	73.2	67.4	538	69.3
FastV [7]	256	16.83(↓36%)	0.92(↑14%)	70.5	72.5	57.9	61.8	86.4	69.3	68.1	70.5	54.2	60.4	431	64.9
VTW [39]	256	15.88(↓39%)	0.96(↑19%)	59.2	66.6	52.8	52.3	77.8	54.3	63.9	57.1	41.0	51.7	338	55.5
Pdrop [67]	256	16.49(↓37%)	0.98(↑21%)	70.6	72.6	57.8	59.5	86.3	69.3	67.8	68.5	50.4	56.1	392	63.5
SparseVLM [78]	256	17.89(↓32%)	0.89(↑10%)	70.7	72.2	58.2	59.8	86.5	70.2	67.9	70.4	51.9	56.6	383	63.9
PixelShuffle [9]	256	16.31(↓38%)	1.02(↑26%)	69.4	69.7	55.0	62.0	87.7	71.4	68.2	66.9	56.4	60.6	402	64.3
DeCo [69]	256	16.27(↓38%)	1.08(↑33%)	70.1	70.3	54.3	61.9	87.4	71.3	68.0	66.3	56.3	60.3	389	64.1
METEOR (Ours)	242*	13.42(↓49%)	1.18(↑46%)	71.4	73.4	58.6	63.5	87.9	72.8	69.8	71.1	71.4	65.6	533	69.0

Table 2. Comparison with efficient MLLMs methods on 11 benchmarks. EAGLE and our model employ the same training data. [†] implies reproducing the results with officially released code on Ascend 910B. Other compared methods are implemented based on EAGLE.

Method	Vis. Tok.	Knowledge				General					OCR and Chart				
		Avg	SQA	MM MU	AI2D	Avg	MME	MMB	SEED	GQA	Avg	Chart QA	OCR Bench	Text VQA	Doc VQA
MGM-HD	2880	62.0	75.1	37.3	73.5	72.7	1606	72.7	73.2	64.5	62.9	59.1	47.7	70.2	74.6
Cambrian-1	576	65.4	80.4	42.7	73.0	73.1	1547	75.9	74.7	64.6	71.3	73.3	62.4	71.7	77.8
METEOR (Ours)	324*	65.4	77.3	42.2	76.8	73.5	1545	78.4	74.4	64.1	74.6	78.5	62.5	73.8	83.7

Table 3. Results using the same training data and vision encoders as Cambrian-1 and LLM is Llama3-8B.



(a) Acc. vs. Throughput. (b) Ablation on allocation strategies.

Figure 6. (a) Accuracy and throughput of full model, pruning with our strategy, attention or similarity for all layers. (b) Ablation on allocation strategies of token sparsity ratio for each encoder.

and Brave-X5 with fewer retained visual tokens and less training data, demonstrating the superiority of our token pruning strategy over MLP and Q-former for integrating tokens from multi-vision encoders. Besides, by equipped with a more advanced LLM, the performance of our model can be further improved, outperforming EAGLE on 7 out of 11 benchmarks with less than 25% of visual tokens. Moreover, even with 126 visual tokens, our method achieves satisfactory performance and consistently performs better than Mini-Gemini, exhibiting the effectiveness in achieving a superior accuracy-efficiency trade-off.

Comparison with Existing Efficient MLLM Methods. Our basic setup follows EAGLE [58] with the same vision encoders and training data, except that we do not involve the pre-alignment stage for simplicity. Thus, we compare with existing efficient MLLM methods based on EAGLE. All the results are tested on Ascend 910B. Table 2 shows that METEOR could reduce 49% TFLOPS and increase 46% throughput with only a 0.3% performance drop on average, showing obvious advantages over existing methods. LLM-based pruning methods such as FastV, Pdrop and Sparse-

Fusion	Pruning	#Tok.	Knowledge	General	OCR	Avg.
Pre-proj	-	1193	64.4	73.7	64.5	67.5
Post-proj	-	1193	65.5	74.0	65.4	68.3
Post-proj	Separate	576	64.8	73.4	62.4	66.9
Post-proj	Random	576	63.4	72.2	48.5	61.4
Post-proj	MLP	576	63.8	72.5	55.9	64.1
Post-proj	PixelShuffle	576	64.1	72.5	55.1	63.9
Post-proj	Resampler	576	61.1	70.6	49.2	60.3
Post-proj	Ours	576	65.4	74.0	65.5	68.3

Table 4. Ablation study on the fusion strategy, i.e., pre-projection and post-projection fusion, and pruning strategy.

VLM ignore the acceleration of vision encoding process, which also takes up a considerable time as the image resolution increases. Besides, a pre-defined pruning ratio for all datasets fails to achieve a promising trade-off between accuracy and efficiency. Moreover, despite satisfactory performance on general benchmarks, training-based methods Deco and PixelShuffle suffer from a significant performance drop on OCR benchmarks, since they fail to achieve the effective instance-adaptive pruning ratio, where OCR tasks demand much more tokens than general tasks.

Extension with Different Vision Encoders. To validate the generalizability of our method, we adopt another group of vision encoders as SigLIP, ConvNeXt, CLIP and DINOv2 following Cambrian-1 [61]. Table 3 shows that METEOR outperforms MGM-HD and Cambrian-1 with fewer visual tokens, especially outperforming by 11.7% and 3.3% accuracy on OCR tasks, exhibiting the effectiveness of our strategy over spatial vision aggregator [61] for adaptively compressing visual tokens from multi-vision encoders.

4.3. Ablation Study

Experiments in this subsection are with 558K data for pre-training and 1M data for SFT. Evaluation results are sum-

Pruning Strategy	Attn-head Filtering	Avg. Tokens	Pruning Layer Indexes	Avg. Retained Tokens	Knowledge	General	OCR	Avg.
Upper Bound	-	576	-	-	65.4	74.0	65.5	68.3
Pre-defined Ratio	✗	297	[1]	[288]	65.0	73.2	59.1	65.8
	✓	297	[1]	[288]	65.1	73.3	59.9	66.1
	✓	242	[4, 12, 20]	[390, 172, 78]	65.0	73.3	62.4	66.9
Our adaptive Ratio	✓	312*	[4, 12, 20]	[525*, 252*, 122*]	65.2	73.8	65.6	68.2
		242*		[396*, 170*, 76*]	65.2	73.5	65.0	67.9
		126*		[226*, 80*, 36*]	64.8	73.4	63.4	67.2

Table 5. Comparison of pre-defined pruning ratio and our instance-adaptive ratio, w or w/o top- k attention head filtering under the same training-free setting. Pruning Layer Indexes specify the LLM layer indexes at which visual tokens are pruned, starting from 0 and occurring prior to input. Token with (*) means the retained token count is not pre-defined and is calculated by averaging across all benchmarks.

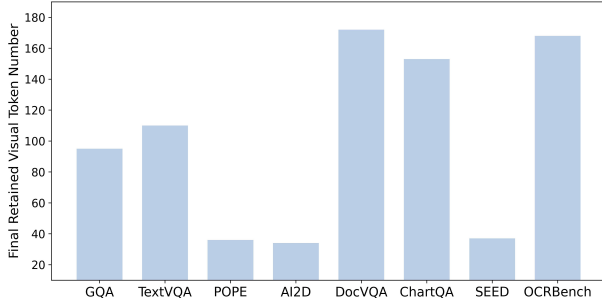


Figure 7. Final retained visual token number for different datasets.

marized into Knowledge (SQA, AI2D, OKVQA), General (GQA, POPE, SEED) and OCR (TextVQA, DocVQA, ChartQA, OCRBench) by averaging to save space.

Token Pruning in Multi-vision Encoding. To identify redundant tokens, we employ the similarity to the average token for the shallow block and attention value in the deep blocks. Figure 6(a) validates the effectiveness of our strategy over using similarity or attention value for all blocks. Besides, to allocate the token sparsity ratio for each encoder, we compare our strategy with 1) Average: equal token numbers to all encoders and 2) Rank-inverse: fewer tokens to encoders generating higher-ranked features. Figure 6(b) shows our rank-based strategy achieves the best performance, especially on more challenging OCR tasks, validating that low-rank feature maps contain less information and should be allocated with fewer token numbers.

Token Pruning in Multi-vision Fusion. Previous methods usually adopt pre-projection fusion, while we propose a more flexible post-projection fusion strategy to independently adapt visual tokens of each encoder before fusion, which achieves superior performance in Table 4. Moreover, based on the projector for aligning multi-encoders, we compare our pruning strategy with 1) Separate pruning: pruning within each encoder separately and 2) Random pruning. Table 4 shows that our strategy significantly outperforms two alternatives, showing the effectiveness of discarding the token redundancy across multi-vision encoders, which also exhibits more feature diversity in Figure 4(b). Our strategy also performs better than parameter-based compression like

MLP and Resampler, which require more data for training.

Instance-adaptive Text-guided Token Pruning. Firstly, Table 5 shows our attn-head filtering strategy significantly improves the performance on OCR by 0.8% accuracy, exhibiting the effectiveness of employing the most relevant attention for accurately measuring the token redundancy. Besides, As shown in Table 5, our strategy consistently outperforms pre-defined fixed pruning rates across all benchmarks with the same average tokens of 242, particularly outperforming by 2.6% on OCR tasks, which are more complex and require more tokens than general tasks. These results demonstrate that tailoring adaptive pruning strategies to specific instances or tasks for dynamically adjusting the token budget could mitigate performance degradation.

Analysis of Retained Token Number. Figure 7 shows the retained token numbers (pruning ratio) vary a lot across datasets. General tasks like AI2D achieve satisfactory results with few tokens, while OCR recognition requiring detailed visual information benefits from retaining more tokens. This adaptive behavior shows the effectiveness of our method in tailoring token budget for varying task demands.

Contribution of Each Encoder. We count the ratio of retained tokens from each encoder in supplementary material.

5. Conclusion

In this paper, we propose a multi-encoder collaborative token pruning strategy, namely METEOR, to progressively eliminate redundant visual tokens across the encoding, fusion and decoding stages. Based on several interesting findings, we leverage the reliable measure to identify redundant tokens across various stages. Besides, the rank of feature maps is employed as a mathematically grounded measure to allocate the sparse ratio for various encoders. Moreover, we propose an instance-adaptive token pruning strategy to dynamically adjust the pruning ratio for various task demands. Extensive experiments demonstrate the effectiveness of our method. Compared with EAGLE, a typical multi-encoder MLLMs, METEOR reduces 76% visual tokens with only 0.3% performance drop, significantly outperforming existing efficient methods.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62431017, Grant 62320106003, Grant U24A20251, Grant 62125109, Grant 62371288, Grant 62301299, Grant 62401357, Grant 62401366, Grant 62120106007, and in part by the Program of Shanghai Science and Technology Innovation Project under Grant 24BC3200800.

References

- [1] Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. HiRED: Attention-guided token dropping for efficient inference of high-resolution vision-language models. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, pages 1773–1781, 2025. 3
- [2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 1, 3, 6
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems* 33, pages 1877–1901, 2020. 1
- [5] Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 3
- [6] Lin Chen and Long Xing. Open-LLaVA-NeXT: An open-source implementation of LLaVA-NeXT series for facilitating the large multi-modal model community. <https://github.com/xiaoachen98/Open-LLaVA-NeXT>, 2024. 6
- [7] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, pages 19–35, 2024. 1, 3, 5, 6, 7
- [8] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67 (12):220101, 2024. 3
- [9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Intern VL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, 2024. 1, 3, 7
- [10] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. <https://lmsys.org/blog/2023-03-30-vicuna/>, 2023. 6
- [11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems* 36, pages 49250–49267, 2023. 6
- [12] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988, 2019. 1
- [13] Shuangrui Ding, Peisen Zhao, Xiaopeng Zhang, Rui Qian, Hongkai Xiong, and Qi Tian. Prune spatio-temporal tokens by semantic-aware temporal accumulation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16945–16956, 2023. 1
- [14] Xiaoran Fan, Tao Ji, Changhao Jiang, Shuo Li, Senjie Jin, Sirui Song, Junke Wang, Boyang Hong, Lu Chen, Guodong Zheng, Ming Zhang, Caishuang Huang, Rui Zheng, Zhiheng Xi, Yuhao Zhou, Shihan Dou, Junjie Ye, Hang Yan, Tao Gui, Qi Zhang, Xipeng Qiu, Xuanjing Huang, Zuxuan Wu, and Yu-Gang Jiang. MouSi: Poly-visual-expert vision-language models. *arXiv preprint arXiv:2401.17221*, 2024. 2, 3, 6
- [15] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 6
- [16] Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang.

- Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, pages 390–406, 2024. 1, 3
- [17] Yuhang Han, Xuyang Liu, Pengxiang Ding, Donglin Wang, Honggang Chen, Qingsen Yan, and Siteng Huang. Rethinking token reduction in MLLMs: Towards a unified paradigm for training-free acceleration. *arXiv preprint arXiv:2411.17686*, 2024. 3
- [18] Xin He, Longhui Wei, Lingxi Xie, and Qi Tian. Incorporating visual experts to resolve the information loss in multimodal large language models. *arXiv preprint arXiv:2401.03105*, 2024. 3
- [19] Yefei He, Feng Chen, Jing Liu, Wenqi Shao, Hong Zhou, Kaipeng Zhang, and Bohan Zhuang. ZipVL: Efficient large vision-language models with dynamic token sparsification and KV cache compression. *arXiv preprint arXiv:2410.08584*, 2024. 3
- [20] Lianyu Hu, Fanhua Shang, Liang Wan, and Wei Feng. iLLaVA: An image is worth fewer than 1/3 input tokens in large multimodal models. *arXiv preprint arXiv:2412.06263*, 2024. 1, 3, 5
- [21] Wenbo Hu, Zi-Yi Dou, Liunan Harold Li, Amita Kamath, Nanyun Peng, and Kai-Wei Chang. Matryoshka query Transformer for large vision-language models. In *Advances in Neural Information Processing Systems 37*, pages 50168–50188, 2024. 3
- [22] Kai Huang, Hao Zou, Ye Xi, BoChen Wang, Zhen Xie, and Liang Yu. IVTP: Instruction-guided visual token pruning for large vision-language models. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, pages 214–230, 2024. 3
- [23] Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6700–6709, 2019. 6
- [24] Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin’e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. From CLIP to DINO: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*, 2023. 2, 3
- [25] Lei Jiang, Weizhe Huang, Tongxuan Liu, Yuting Zeng, Jing Li, Lechao Cheng, and Xiaohua Xu. FoPru: Focal pruning for efficient large vision-language models. *arXiv preprint arXiv:2411.14164*, 2024. 1, 3, 4
- [26] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25004–25014, 2025. 5
- [27] Oğuzhan Fatih Kar, Alessio Tonioni, Petra Poklukar, Achin Kulshrestha, Amir Zamir, and Federico Tombari. BRAVE: Broadening the visual encoding of vision-language models. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, pages 113–132, 2024. 2, 3, 4, 6
- [28] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Proceedings of the 14th European Conference on Computer Vision (ECCV)*, pages 235–251, 2016. 6
- [29] Byung-Kwan Lee, Beomchan Park, Chae Won Kim, and Yong Man Ro. MoAI: Mixture of all intelligence for large language and vision models. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, pages 273–302, 2024. 3
- [30] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 18893–18912, 2023. 6
- [31] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. SEED-Bench: Benchmarking multimodal large language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13299–13308, 2024. 6
- [32] Hongliang Li, Jiaxin Zhang, Wenhui Liao, Dezhi Peng, Kai Ding, and Lianwen Jin. RedundancyLens: Revealing and exploiting visual token processing redundancy for efficient decoder-only MLLMs. *arXiv preprint arXiv:2501.19036*, 2025. 3
- [33] Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. TokenPacker: Efficient visual projector for multimodal LLM. *International Journal of Computer Vision*, pages 1–19, 2025. 3, 6
- [34] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, 2023. 6
- [35] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-Gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024. 1, 2, 3, 6
- [36] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26763–26773, 2024. 1, 2, 3
- [37] Mingbao Lin, Rongrong Ji, Yan Wang, Yichen Zhang, Baochang Zhang, Yonghong Tian, and Ling Shao. HRank: Filter pruning using high-rank feature map. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1529–1538, 2020. 4
- [38] Ziyi Lin, Dongyang Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Yu Qiao, and Hongsheng Li. SPHINX: A mixer of weights, visual embeddings and image scales for multi-modal large language

- models. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, pages 36–55, 2024. 3, 6
- [39] Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, pages 5334–5342, 2025. 7
- [40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36*, pages 34892–34916, 2023. 1, 3
- [41] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024. 6
- [42] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. <https://llava-vl.github.io/blog/2024-01-30-llava-next/>, 2024. 2, 3, 6
- [43] Shikun Liu, Linxi Fan, Edward Johns, Zhiding Yu, Chaowei Xiao, and Anima Anandkumar. Prism: A vision-language model with multi-task experts. *Transactions on Machine Learning Research*, 2024. 2, 3
- [44] Ting Liu, Liangtao Shi, Richang Hong, Yue Hu, Qunjun Yin, and Linfeng Zhang. Multi-stage vision token dropping: Towards efficient multimodal large language model. *arXiv preprint arXiv:2411.10803*, 2024. 3
- [45] Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of OCR in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023. 6
- [46] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, pages 216–233, 2024. 6
- [47] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022. 6
- [48] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. DeepSeek-VL: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024. 3
- [49] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems 35*, pages 2507–2521, 2022. 6
- [50] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 4, 6
- [51] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3190–3199, 2019. 6
- [52] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, 2022. 6
- [53] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. DocVQA: A dataset for VQA on document images. In *2021 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2200–2209, 2021. 6
- [54] Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruvi Shah, Xi-anzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu He, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Gräsch, Alexander Toshev, and Yinfei Yang. MM1: methods, analysis and insights from multimodal LLM pre-training. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, pages 304–323, 2024. 6
- [55] Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. DeepStack: Deeply stacking visual tokens is surprisingly simple and effective for LMMs. In *Advances in Neural Information Processing Systems 37*, pages 23464–23487, 2024. 6
- [56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 2, 3, 6
- [57] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PrüMerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024. 1, 3, 4
- [58] Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, and Guilin Liu. Eagle: Exploring the design space for multimodal LLMs with mixture of encoders. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 2, 3, 4, 6, 7
- [59] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8317–8326, 2019. 6
- [60] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue

- Cao. EVA-CLIP: Improved training techniques for CLIP at scale. *arXiv preprint arXiv:2303.15389*, 2023. 3, 6
- [61] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems 37*, pages 87310–87356, 2024. 1, 2, 3, 7
- [62] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9568–9578, 2024. 3
- [63] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 1, 6
- [64] Ao Wang, Fengyuan Sun, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. [CLS] token tells everything needed for training-free efficient MLLMs. *arXiv preprint arXiv:2412.05819*, 2024. 3, 4
- [65] Haicheng Wang, Zhemeng Yu, Gabriele Spadaro, Chen Ju, Victor Quéru, and Enzo Tartaglione. FOLDER: Accelerating multi-modal large language models with enhanced performance. *arXiv preprint arXiv:2501.02430*, 2025. 3
- [66] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language model. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, pages 408–424, 2024. 2
- [67] Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. PyramidDrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*, 2024. 1, 3, 5, 6, 7
- [68] Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Mitigating hallucination in large vision-language models via modular attribution and intervention. In *The Thirteenth International Conference on Learning Representations*, 2025. 5
- [69] Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. DeCo: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:2405.20985*, 2024. 1, 3, 7
- [70] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mPLUG-Owl2: Revolutionizing multi-modal large language model with modality collaboration. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13040–13051, 2024. 6
- [71] Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. Fit and prune: Fast and training-free visual token pruning for multimodal large language models. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, pages 22128–22136, 2025. 1, 3, 5
- [72] Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. ATP-LLaVA: Adaptive token pruning for large vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24972–24982, 2025. 3, 5
- [73] Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, Ying Shan, and Yansong Tang. VoCo-LLaMA: Towards vision compression with large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29836–29846, 2025. 3
- [74] Hao Yin, Guangzong Si, and Zilei Wang. Lifting the veil on visual information flow in MLLMs: Unlocking pathways to faster inference. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9382–9391, 2025. 1, 3
- [75] Qizhe Zhang, Aosong Cheng, Ming Lu, Zhiyong Zhuo, Minqi Wang, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. [CLS] attention is all you need for training-free visual token pruning: Make VLM inference faster. *arXiv preprint arXiv:2412.01818*, 2024. 1, 3, 4
- [76] Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. LLaVA-Mini: Efficient image and video large multimodal models with one vision token. In *The Thirteenth International Conference on Learning Representations*, 2025. 3
- [77] Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in LVLMs. *arXiv preprint arXiv:2411.09968*, 2024. 5
- [78] Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. SparseVLM: Visual token sparsification for efficient vision-language model inference. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. 1, 3, 5, 7
- [79] Zeliang Zhang, Phu Pham, Wentian Zhao, Kun Wan, Yu-Jhe Li, Jianing Zhou, Daniel Miranda, Ajinkya Kale, and Chenliang Xu. Treat visual tokens as text? but your MLLM only needs fewer efforts to see. *arXiv preprint arXiv:2410.06169*, 2024. 3
- [80] Shiyu Zhao, Zhenting Wang, Felix Juefei-Xu, Xide Xia, Miao Liu, Xiaofang Wang, Mingfu Liang, Ning Zhang, Dimitris N Metaxas, and Licheng Yu. Accelerating multimodal large language models by searching optimal vision token reduction. *arXiv preprint arXiv:2412.00556*, 2024. 3
- [81] Yiwu Zhong, Zhuoming Liu, Yin Li, and Liwei Wang. AIM: Adaptive inference of multi-modal LLMs via token merging and pruning. *arXiv preprint arXiv:2412.03248*, 2024. 3
- [82] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 3
- [83] Yuke Zhu, Chi Xie, Shuang Liang, Bo Zheng, and Sheng Guo. FocusLLaVA: A coarse-to-fine approach for effi-

cient and effective visual token compression. *arXiv preprint arXiv:2411.14228*, 2024. [3](#)

- [84] Zhuofan Zong, Bingqi Ma, Dazhong Shen, Guanglu Song, Hao Shao, Dongzhi Jiang, Hongsheng Li, and Yu Liu. MoVA: Adapting mixture of vision experts to multimodal context. In *Advances in Neural Information Processing Systems* 37, pages 103305–103333, 2024. [2](#), [3](#)