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

Yuchen Liu[1], Yaoming Wang[3†], Bowen Shi[1], Xiaopeng Zhang[2†], Wenrui Dai[1*], Chenglin Li[1], Hongkai Xiong[1] and Qi Tian[2]

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

{liuyuchen6666, sjtu_shibowen, daiwenrui, lcl1985, xionghongkai}@sjtu.edu.cn
wangyaoming03@meituan.com, zxphistory@gmail.com, tian.qi1@huawei.com

A. Diversity Measurement

For a given token sequence $A \in \mathbb{R}^{N \times D}$ with the number of tokens as N and the dimension as D, we analyze the token representations on the unit sphere by normalizing the length of each token to one following the common practice [3, 22]. As the convex envelope of matrix rank, the nuclear norm is widely used to measure the diversity of matrix rows [2]. Based on the theorem in [4], $\|A\|_*$ is the convex envelope to $rank(A)$ when $\|A\|_F \leq 1$. Since $\|A\|_F = \sqrt{N}$ (each row vector in A is unit vector), the convex envelope of $rank(A)$ turn to $\|A\|_*/\sqrt{N}$. Due to the properties of the nuclear norm, $\|A\|_*/\sqrt{N}$ is bounded by $\sqrt{k}$ where $k = \min(N, D)$. In practical implementation, the length of the token sequence N is typically around 1024, and the embedding dimension D is commonly 4096. Thus, $\min(N, D) = N$ and $\|A\|_*/\sqrt{N} \leq \sqrt{N}$. Since the length of token sequences varies during the forward process, the upper bound of $\|A\|_*/\sqrt{N}$ depends on N. To get rid of the influence from N, we apply $\|A\|_*/N$, which has a clear upper bound to measure the diversity in token sequences.

B. Experimental Settings

B.1 Implementation Details

We train the model for one epoch, and all experiments are conducted on 8 Ascend 910B GPUs with 65 GB of memory. We randomly sample 0.1% from the training set of the 665k instruction data of LLaVA-1.5 [11] to estimate the feature map rank of different vision encoders. The training hyper-parameters, e.g., batch size, learning rate and weight decay, all follow EAGLE [19]. For multi-vision encoding, we divide the whole encoder equally into three stages, e.g, conducting token pruning after the 7th, 15th and 23rd block for an encoder with 24 blocks in total. For pruning in LLM decoding, we conduct token reduction at the 4th, 12th and 20th layer, and filter with top-5 attention heads for more accurately measuring the redundancy of visual tokens.

B.2 Evaluation Datasets

GQA [6] consists of three components: scene graphs, questions, and images. The image component includes raw images, spatial image features, and features of all objects within the images. GQA questions are designed to assess a model's understanding of visual scenes and its ability to reason about various image attributes.
MMBench [13] provides a multi-dimensional, hierarchical evaluation of a model's overall performance. It structures ability assessments across three levels: the first level (L-1) evaluates two primary abilities, perception and reasoning; the second level (L-2) expands on L-1 with six sub-abilities; and the third level (L-3) further refines L-2 into 20 specific ability dimensions. This tiered structure allows for a thorough evaluation of the diverse capabilities.
MME [5] is a comprehensive benchmark designed to meticulously evaluate a model's performance across various aspects. It comprises 14 sub-tasks that specifically assess both perceptual and cognitive abilities. The use of manually constructed instruction-answer pairs and concise instruction design effectively mitigates data leakage and ensures fair performance evaluation.
POPE [9] evaluates object hallucination in models by reformulating hallucination assessment. Models are tasked with answering binary questions regarding the presence of objects in images. Accuracy, recall, precision, and F1-score are utilized to measure hallucination levels across three distinct sampling strategies.
ScienceQA [14] encompasses a diverse range of scientific domains, including natural, language, and social sciences. Questions are hierarchically organized by topic, category, and skill, resulting in 26 topics, 127 categories, and 379 skills. This structure provides a comprehensive and diverse question set for evaluating a model's multimodal understanding, multi-step reasoning, and interpretability.

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

Strategy	GQA	OKVQA	SEED	SQA	AI2D	POPE	TextVQA	DocVQA	ChartQA	OCR	Avg
Rank-based	64.4	58.8	70.8	69.8	67.8	86.7	71.3	72.2	66.9	513	67.99
Average	64.2	58.3	70.0	69.0	67.9	87.3	69.7	67.6	65.1	496	66.87
Rank-reverse	62.9	57.9	68.8	69.1	67.4	85.9	68.1	65.5	64.1	472	65.69

Table 1. Ablation study on the assigning strategies of the retained visual token number budget for different vision experts.

Strategy	#Token	GQA	OKVQA	SEED	SQA	AI2D	POPE	TextVQA	DocVQA	ChartQA	OCR
Post-projection fusion	1193	64.4	58.8	70.8	69.8	67.8	86.7	71.3	72.2	66.9	513
Pre-projection fusion	1193	64.3	57.9	70.0	69.4	66.0	86.7	71.6	70.9	66.2	493
Fusion strategy ↑; Pruning strategy ↓											
Upper bound	1193	64.4	58.8	70.8	69.8	67.8	86.7	71.3	72.2	66.9	513
Our collaborative prune	576	64.3	58.7	70.7	69.7	67.7	86.9	71.5	72.2	66.3	521
Random prune	576	63.3	57.9	68.4	68.2	64.0	84.9	60.9	43.0	47.0	432
Separate prune	576	63.6	58.6	69.2	69.0	66.8	87.3	68.3	66.2	65.0	502
Resampler	576	61.7	54.3	64.5	66.6	62.3	85.5	63.1	48.1	50.9	348
MLP	576	63.2	58.1	68.0	67.2	66.0	86.3	67.8	54.7	60.1	410
PixShuffle	576	63.6	58.2	67.9	68.2	65.9	85.9	67.2	53.4	59.4	400

Table 2. Ablation study on the pruning strategies for the multi-vision experts fusion stage.

TextVQA [20] evaluates a model's ability to integrate diverse textual information within images. It assesses text understanding and reasoning through visual question-answering tasks that incorporate rich textual data. To answer accurately, models must comprehend both the visual content and the embedded text.

SEED-Bench [8] comprises 19k multiple-choice questions with precise human annotations, approximately six times larger than existing benchmarks. It evaluates 12 dimensions comprehension of image modalities.

ChartQA [16] covers 9.6K human-written questions as well as 23.1K questions generated from human-written chart summaries, which involve visual and logical reasoning over charts.

DocVQA [17] promotes the extraction and utilization of document content to address high-level tasks defined by human users. Challenges and release datasets are organized to enable machines to comprehend document images and answer associated questions.

AI2D [7] comprises over 5k science diagrams from grade school curricula, accompanied by more than 150k detailed annotations, their syntactic parses, and over 15k associated multiple-choice questions.

OKVQA [15] includes outside knowledge visual question answering datasets with more than 14k questions that require external knowledge to answer.

OCRBench [12] is a comprehensive evaluation benchmark for OCR, which contains various text-related visual tasks including Text Recognition, Scene Text-Centric Visual Question Answering (VQA), Document-Oriented VQA, Key Information Extraction (KIE), and Handwritten Mathematical Expression Recognition (HMER) collected from 29 datasets.

C. Additional Experimental Results

Token Pruning in Multi-vision Encoding Ablation study for the assigning strategy of retained visual token number ratio for different vision experts is shown in Table 1. To determine the token number allocation for multi-vision experts, we compare our rank-based strategy with 1) Average: assigning equal token numbers to all experts and 2) Rank-inverse: assigning fewer tokens to experts generating higher-ranked features. Table 1 shows that our rank-based strategy achieves the best performance, especially on more challenging OCR tasks, demonstrating that low-rank feature maps contain less information and should be allocated with fewer token budget.

Token Pruning in Multi-vision Fusion Previous multi-vision experts based MLLMs usually adopt pre-projection fusion, while we propose a more flexible post-projector fusion strategy that each expert independently adapts visual tokens before fusion, which achieves superior performance as shown in Table 2. Moreover, based on the projector for aligning multi-vision experts, we compare our collaborative pruning strategy with 1) Separate pruning: pruning tokens within each expert separately and 2) Random pruning. Table 2 shows that our strategy significantly outperforms two alternatives, showing the effectiveness of reducing the token redundancy across multi-vision experts. Compared with parameter-based compression methods, our parameter-free pruning strategy performs better with the priority of measuring the redundancy with similarity across experts.

Instance-adaptive Text-guided Token Pruning Firstly, we compare the performance of employing all attention heads with selecting top-k significant attention heads for evaluating the visual token redundancy. Table 3 shows that our attn-head filtering strategy significantly improves the per-

Pruning Strategy	Attn-head Filtering	Avg. Tokens	Pruning Layer Indexes	Avg. Retained Tokens	OKVQA	SQA	AI2D	GQA	SEED	POPE	TextVQA	DocVQA	ChartQA	OCR
Upper Bound	-	576	-	-	58.7	69.7	67.7	64.3	70.7	86.9	71.5	72.2	66.3	521
Pre-defined Ratio	✗	297	[1]	[288]	58.8	69.8	66.4	63.4	70.0	86.1	70.0	59.9	63.4	432
	✓	297	[1]	[288]	58.7	69.7	66.8	63.5	70.2	86.2	70.2	61.4	63.2	447
	✓	242	[4, 12, 20]	[390, 172, 78]	59.0	69.5	66.5	63.2	70.6	86.2	70.5	65.9	64.8	485
Task-adaptive Ratio	✓	312^{*}	[4, 12, 20]	[525^{*}, 252^{*}, 122^{*}]	59.0	70.0	66.8	64.0	71.0	86.4	71.3	70.1	65.6	552
		242^{*}		[396^{*}, 170^{*}, 76^{*}]	59.0	69.9	66.7	63.6	70.7	86.2	70.7	69.0	65.1	550
		126^{*}		[226^{*}, 80^{*}, 36^{*}]	58.4	69.7	66.4	63.5	69.8	86.9	70.1	66.7	64.6	523

Table 3. Comparison with pre-defined pruning strategy under the same training setting using common benchmarks. Pruning Layer Indexes specify the LLM layer indexes at which vision tokens are pruned, starting from 0 and occurring prior to input.

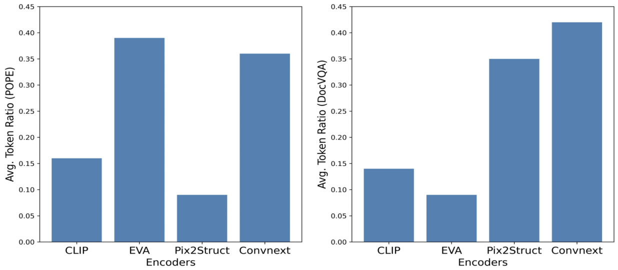

Figure 1. Contribution of vision encoders for POPE and DocVQA.

formance on OCR by 0.8% accuracy, exhibiting the effectiveness of employing the most relevant attention for accurately measuring the token redundancy. To evaluate the adaptability of our instance-adaptive token pruning strategy, we conduct experiments with a pre-defined pruning ratio to prune visual tokens at a fixed ratio. As shown in Table 3, our strategy consistently outperforms 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 illustrate that tailoring adaptive pruning strategies to specific instances can dynamically adjust the token budget and help mitigate performance degradation.

Contribution of Vision Encoders. To understand the impact of vision encoders, we evaluate the ratio of retained token numbers for different vision encoders. As shown in Fig. 1, EVA and Convnext contribute more to fine-grained POPE, while Pix2Struct and Convnext are more crucial for DocVQA, showing the effectiveness of our method for retaining suitable visual tokens for different tasks.

Extensive Evaluation on Video Benchmarks. Moreover, we make zero-shot evaluation on video benchmarks following IG-VLM by directly using our image MLLMs. Table 4 shows that METEOR achieves superior or comparable results with fewer tokens. These results support the generality of METEOR for token compression in various benchmarks.

Method	# Tok.	Next-QA	MSVD		ActivityNet	
		Acc.	Acc.	Score	Acc.	Score
Video-LLaVA	2048	-	70.7	-	45.3	-
IG-VLM	2880	63.1	78.8	4.1	54.3	3.4
EAGLE	1024	63.7	79.2	4.1	55.0	3.4
DeepStack-L	576	61.0	76.0	4.0	49.3	3.1
METEOR	242*	63.3	79.0	4.1	55.5	3.4

Table 4. Zero-shot evaluation on Video QA benchmarks.

Method	#Tok.	*Knowledge*	*General*	*OCR and Chart*
MGM-HD [10]	2880	62.0	72.7	62.9
Cambrian-1 [21]	576	65.4	73.1	71.3
METEOR	576*	65.7$_{\pm0.1}$	73.8$_{\pm0.3}$	75.6$_{\pm0.3}$
METEOR	324*	65.5$_{\pm0.1}$	73.5$_{\pm0.2}$	74.6$_{\pm0.3}$

Table 5. Results using the same training data, vision encoders and LLM as Cambrian-1, averaged over three tests with std reported.

Extension on Different Set of Vision Encoders. Table 5 shows that METEOR consistently outperforms Cambrian-1 on *Knowledge* and *General* tasks that rely more on LLM reasoning, and *OCR and Chart* for fine-grained visual reasoning using the same 576 tokens. Remarkably, the performance gains of METEOR become significant on *OCR and Chart* (4.3% using 576 tokens and 3.3% using 324 tokens), since the task demands handling detailed local visual information. These results strongly support the efficacy of METEOR's advanced token pruning.

Extensive Evaluation on Single-encoder MLLM. We validate the effectiveness of our pruning strategies at each stage for single-encoder MLLMs. Table 6 shows that our progressive pruning in *Stage 1* outperforms pruning all at once in PruMerge++ on 5 of 6 benchmarks following the setting of PruMerge. Table 7 shows that our adaptive pruning in *Stage 3* outperforms FastV and Pdrop on tasks with varying complexity, especially the challenging OCR tasks (61.8% vs. 55.9% and 59.1% for ChartQA and 67.5% vs. 62.1% and 65.6% for DocVQA). Furthermore, the combination of *Stages 1&3* yields a highly efficient LLaVA variant of 12% FLOPS and is superior to AIM [24] (to be cited in the final version) on most benchmarks, especially on TextVQA (53.8% vs. 48.4%). This forms the solid basis for METEOR to achieve efficient multi-encoder MLLM by using the feature rank for sparsity allocation and pruning mutually redundant tokens in multi-vision fusion.

Ablation on Hyperparameters. Table 8 shows that selecting top-5 (k=5) attention heads to identify the redundant visual tokens performs best, and λ controls the different over-

Method	Flops(%)	GQA	SQA	MME	POPE	TextVQA	VQAv2
LLaVA-1.5-7b	100%	62.0	66.8	1510.7	85.9	58.2	78.5
+PruMerge++ [18]	29%	57.5	68.3	1462.4	84.0	57.1	76.8
+Our Stage 1	29%	58.6	69.1	1472.8	84.5	57.3	76.4
+AIM [24]	12%	54.6	67.1	1277.7	79.5	48.4	69.0
+Our Stage 1&3	12%	55.1	67.9	1321.2	78.4	53.8	69.5

Table 6. Comparison of each component based on LLaVA-1.5.

Method	Flops(%)	GQA	SQA	POPE	TextVQA	ChartQA	DocVQA
LLaVA-Next-7b	100%	64.2	70.4	86.1	67.2	64.0	70.0
+FastV [1]	51%	63.5	69.3	86.3	66.5	55.9	62.1
+Pdrop [23]	46%	63.9	69.4	86.4	67.0	59.1	65.6
+Our Stage 3	44%	64.0	69.6	86.1	67.1	61.8	67.5

Table 7. Comparison of pruning methods within LLM layers.

k	# Tok.	*Know.*	*Gene.*	*OCR*	λ	# Tok.	*Know.*	*Gene.*	*OCR*
3	242*	65.0	73.6	64.8	65	312*	65.2	73.8	65.6
5	242*	65.2	73.5	65.0	50	242*	65.2	73.5	65.0
7	242*	64.9	73.5	64.6	25	126*	64.8	73.4	63.4

Table 8. Ablation study on hyperparameters k and λ.

all token budgets for the model.

D. Limitation

While we have validated that METEOR performs well across a wide range of benchmarks, we find that it lags behind the best models available, such as Qwen2-VL. One reason is that our training data is not diverse and huge enough. Besides, a fixed input resolution may suffer poor performance for the given image with an extreme imbalanced ratio. In the future, we will validate our method based on more advanced architecture equipped with anyres techniques to accommodate images of various high resolutions with more diverse data for improved performance.

References

[1] 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. 4

[2] Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3941–3950, 2020. 1

[3] Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 3015–3024, 2021. 1

[4] Maryam Fazel. *Matrix rank minimization with applications*. PhD thesis, PhD thesis, Stanford University, 2002. 1

[5] 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. 1

[6] 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. 1

[7] 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. 2

[8] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 2

[9] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 1

[10] 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. 3

[11] 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. 1

[12] 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. 2

[13] 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. 1

[14] 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. 1

[15] 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 3195–3204, 2019. 2

[16] 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. *arXiv preprint arXiv:2203.10244*, 2022. 2

[17] 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. 2

[18] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*, 2024. 4

[19] 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. *arXiv preprint arXiv:2408.15998*, 2024. 1

[20] 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. 2

[21] 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. 3

[22] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 9929–9939, 2020. 1

[23] 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. 4

[24] 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, 4