DivPrune: Diversity-based Visual Token Pruning for Large Multimodal Models Supplementary Material

Saeed Ranjbar Alvar, Gursimran Singh†, Mohammad Akbari†, Yong Zhang
Huawei Technologies Canada Co., Ltd.
{saeed.ranjbar.alvar1, gursimran.singh1, mohammad.akbari, yong.zhang3}@huawei.com

1. Datasets, Tasks, and Metrics

We briefly introduce the 11 image-language and 5 video-language datasets used in the experiments of the main manuscript. In addition, the system prompt (instruction) used to get output results for each dataset is given. The details of datasets used for image-language and video-language understanding tasks are presented in Tab. 2. Furthermore, the details on 3 extra datasets used for our new experiments in the supplementary material are provided.

As shown in the table, diverse range of tasks including image captioning, visual reasoning, open-ended visual question answering, closed-ended visual question answering, and multiple-choice visual question answering are used to evaluate the performance of the visual token pruning methods compared with ours. Note that the system prompts are the default prompts provided in the lmms-evals evaluation package [24].

2. More Examples for Insights

In Fig. 3 of the main manuscript, DivPrune and an importance-based token pruning method (i.e., FastV [3]) are compared using (a) t-SNE visualization for a sample input's visual tokens and (b) a histogram of the max-min distance between the selected tokens across 1000 data samples from SeedBench dataset [8]. In this section, additional examples from SeedBench and GQA datasets [7] are respectively provided in Fig. 1-(a)-(b) and Fig. 1-(c)-(f).

As shown in Fig. 1-(a)-(b), similar to the observation in the main manuscript, the majority of the selected tokens using FastV method are densely clustered near each other, whereas the tokens selected using DivPrune are more widely separated. As a result, the redundancy among the selected tokens decreases. In addition, unlike DivPrune, FastV does not include any tokens from the top clusters. Hence, DivPrune achieve a better representation for the original set of tokens.

Further examples using GQA dataset are provided in Fig. 1-(a)-(e). Inline with earlier observation, Divprune reduces redundancy and achieves better representation compared to importance-based token pruning when applied to GQA dataset. To verify this behavior over multiple dataset samples, the max-min distance among the selected visual tokens is obtained using 1000 randomly selected samples from the GQA. The histogram of the obtained max-min values for DivPrune and FastV is shown Fig. 1-(f). The histogram also verifies that our method achieves higher max-min distance values, thereby reducing redundancy for the tested samples of the dataset.

3. Results with Additional Datasets

In addition to the datasets tested in the main manuscript, we evaluate the proposed method and the baselines with LLaVA 1.5-7b model on more visual question answering datasets: TextVQA [19], VizWiz [6], and VQAv2 [5]. The details corresponding to each dataset are included in Tab. 2. The same hyperparameters used for results in Tab. 1 of the main manuscript are applied to both our method and the baselines. The results for the proposed method and the baselines are summarized in Tab. 1. The TFLOPs are calculated for each dataset, and the average TFLOP and ratio are given in the TFLOP column of the table. VTW, FastV, and ours are the 3 training-free and calibration-free methods. As the results indicate, our method outperforms VTW

	Method	TFLOP (ratio %)	TextVQA EM	VizWiz EM	VQAv2 EM
LLaVA 1.5-7B	Original	3.13 (100.00)	46.08	54.24	76.65
	VTW [11]	0.507 (16.20)	8.22	50.13	42.13
	FastV [3]	0.418 (13.35)	8.21	50.48	41.71
	Ours	0.416 (13.29)	**35.97**	**57.41**	**71.55**
	PruMerge [18]	Variable	**37.70**	56.31	65.01
	Ours*	Variable	35.00	**57.43**	**69.59**
	FitPrune$^{\triangle}$ [21]	0.417 (13.32)	30.10	54.62	64.86
	M$^{3\bullet}$ [2]	0.416 (13.29)	44.31	52.98	75.87

Table 1. Comparison results of our method and baselines on three additional datasets. $\bullet$: Finetuning is used, $\triangle$: Calibration dataset is used. **Ours***: Our method matching the PruMerge selection ratio.

	Dataset	Task	Metric	System Prompt
Image-Language Understanding	COCO-2017 [10]	Image Captioning	CIDEr	Provide a one-sentence caption for the provided image.
	Flicker30k [17]	Image Captioning	CIDEr	Provide a one-sentence caption for the provided image.
	GQA [7]	CE-VQA	Eaxct Match	Answer the question using a single word or phrase.
	MMBench [12]	MC-VQA	Accuracy	Answer with the option's letter from the given choices directly.
	MME [4]	CE-VQA	Perception Score	Answer the question using a single word or phrase.
	MMU [23]	CE-VQA and OE-VQA	Accuracy	Answer with the option's letter from the given choices directly, OR Answer the question using a single word or phrase.
	Nocaps [1]	Image Captioning	CIDEr	Provide a one-sentence caption for the provided image
	OKVQA [16]	Visual Reasoning	Exact Match	When the provided information is insufficient, respond with 'Unanswerable'. Answer the question using a single word or phrase.
	POPE [9]	CE-VQA	F1 Score	Answer the question using a single word or phrase.
	ScienceQA-Image [13]	Visual reasoning	Exact Match	Answer with the option's letter from the given choices directly.
	SeedBench-Image [8]	MC-VQA	Accuracy	Answer with the option's letter from the given choices directly.
	TextVQA [19]	CE-VQA	Exact Match	Answer the question using a single word or phrase.
	VizWiz [6]	CE-VQA	Exact Match	When the provided information is insufficient, respond with 'Unanswerable'. Answer the question using a single word or phrase.
	VQAv2 [5]	CE-VQA	Exact Match	Answer the question using a single word or phrase.
Video-Language	ActivityNet [22]	CE-VQA	Accuracy/ GPT-Assisted score	Answer the question using a single word or phrase.
	SeedBench-Video [8]	MC-VQA	Accuracy	Answer with the option's letter from the given choices directly.
	VideoChatGPT-temporal [14]	OE-VQA	GPT-Assisted-score	Evaluate the temporal accuracy of the prediction compared to the answer.*
	NextQA [20]	CE-VQA	WUPS	Answer a question using a short phrase or sentence.
	EgoSchema [15]	MC-VQA	Accuracy	Answer with the option's letter from the given choices directly.

Table 2. Details of the datasets, the corresponding tasks, metrics, and prompts used in our experiments. CE-VQA: Closed-Ended Visual Question Answering, OE-VQA: Open-Ended Visual Question Answering, MC-VQA: Multiple-Choice Visual Question Answering. *: Only the main sentence from the prompt is shown here.

and FastV on TextVQA, VizWiz, and VQAv2 datasets by $\approx$ 27%, 7%, and 29%, respectively.

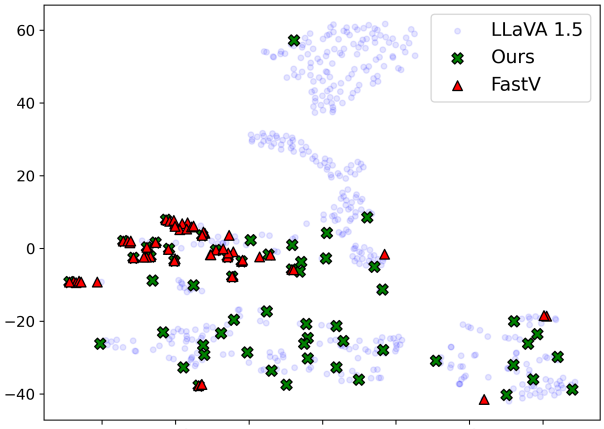
In the case of dynamic pruning scenario, we matched the pruning ratio with that of the PruMerge baseline [18]. The comparison of our results with PruMerge reveals that our method achieves higher accuracy on VizWiz and VQAv2 datasets. Compared to FitPrune [21], which uses calibration datasets to optimize the procedure of token pruning, we achieve higher task performance on all the datasets. Finally, compared to the fine-tuning-based M^3 [2] method, our performance is worse on TextVQA, comparable on VQAv2, and better on VizWiz dataset. DivPrune achieves better results compared to the original model on VizWiz dataset. Visual token pruning has been shown to improve the original model's performance for some datasets [3]. Overall, the results shown in the table are inline with the results reported in the manuscript. This proves that DivPrune outperforms baselines on a diverse range of tasks and datasets.

3.1. Different TFLOPs for the 13b Model

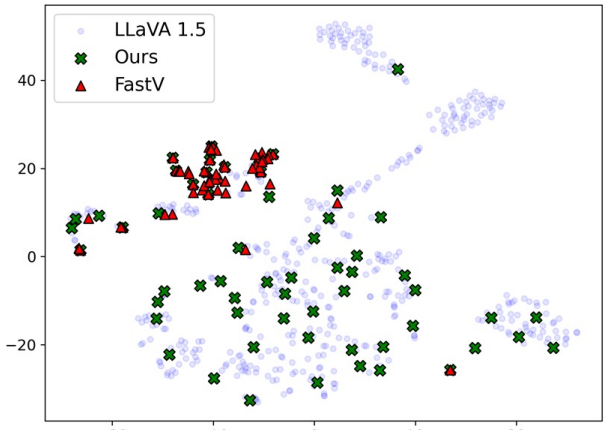
In the main manuscript, we showed the performance of baselines and our method across various TFLOP ratios for LLaVA 1.5-7b model. In this section, we present the results with LLaVA 1.5-13b model. The results are shown in Fig. 2 where the y-axis represents average performance on four datasets, namely, COCO (CIDEr), OKVQA (Acc), POPE (F1), and MMBench (Acc). For all datasets, the performance metric spans from 0 to 1, with the exception of the CIDEr metric, which can reach a peak value of 1.16 for the tested model. On the x-axis, we only show the high compression scenario (TFLOP ratio $\leq 40\%$). As shown in the figure, our method significantly outperforms all the baselines, particularly in high compression scenarios (TFLOP $\leq 25\%$). Furthermore, the gap between our approach and the baselines increases at extreme compression levels. For higher TFLOP ratios almost all methods converge toward the original performance. The pruning ratio and calibration samples for the FitPrune are not provided for the 13b model, unlike the 7b model, hence it is excluded from the baselines.

4. Qualitative Results

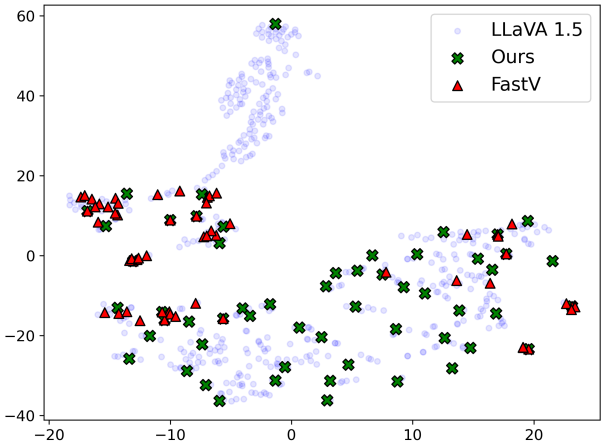
In this section, we present some qualitative results comparing the proposed method with the relevant baselines. Given the significant improvement of our method over the baselines on image captioning tasks, we provide 3 examples for image captioning using COCO [10] dataset in Fig. 3. For all the examples, the prompt, ground truth (GT) caption, and the LLaVA 1.5-7B model's output are given for reference. The model's output when our pruning method and baselines are applied is also shown for each example. We follow the experimental settings used to obtain the results in Tab. 1 of the main manuscript. The results show that using DivPrune (our method) enables the model to produce descriptions that closely align with the original model's output, which is very similar to the ground truth, while only



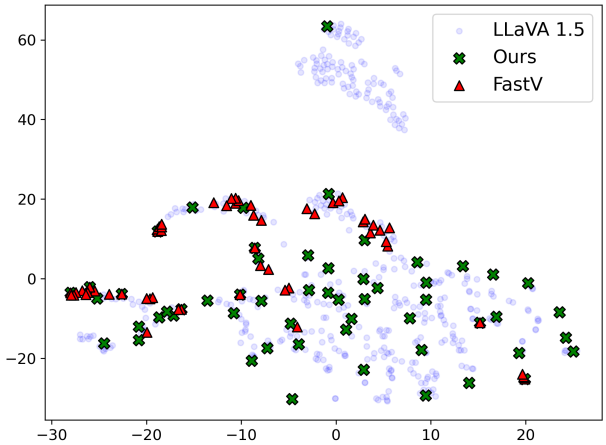
(a) SeedBench example #1



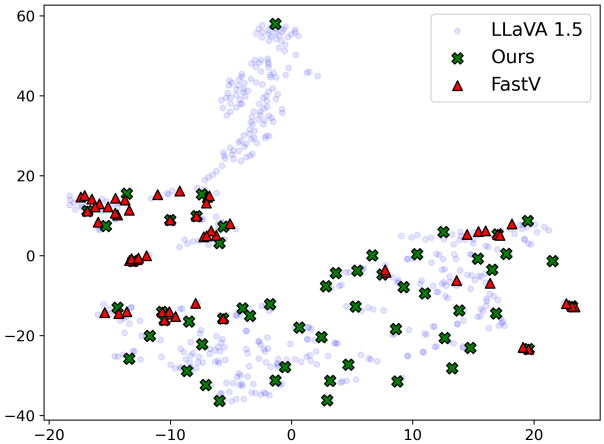
(b) SeedBench example #2



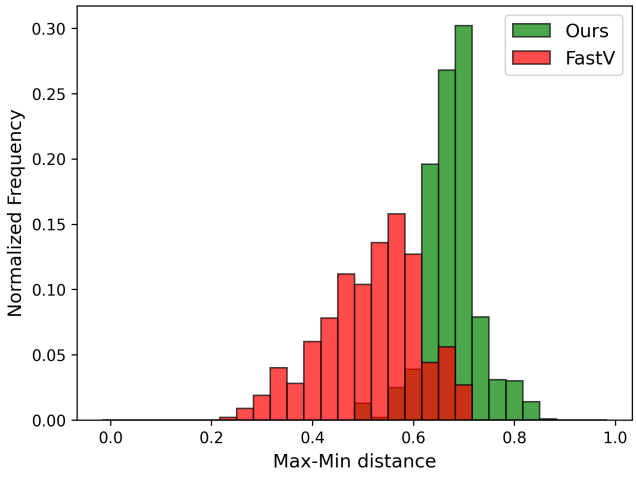
(c) GQA example #1



(d) GQA example #2



(e) GQA example #3



(f) Histogram of Max-Min Distance for GQA dataset

Figure 1. (a)-(b) t-SNE visualization of visual tokens using SeedBench samples, (c)-(e) t-SNE visualization of visual tokens using GQA samples, (f) Histogram of the Max-Min distance between the selected tokens over the GQA dataset.

using 12% TFLOP compared to the original model. In contrast, FastV and VTW generate irrelevant captions for the given images with the same TFLOP ratio.

We also provide qualitative examples for a visual question answering task. Specifically, the output of LLaVA 1.5-7B model for sample images and questions from OKVQA [16] dataset, along with the ground truth and the corresponding prompt are shown in Fig. 4. As the figure illustrates, the output of the model when DivPrune is applied matches the ground truth. However, when FastV or VTW method are used, the model either generates incorrect responses or indicates that insufficient information is provided.

5. Hyper-Parameters of Baselines

In the main manuscript, TFLOP ratio and values are reported for DivPrune and the baselines. In this section, we provide the details on the hyperparameters specific to these methods. For DivPrune, the pruning ratio is set to 91.2%. For FastV with 7B models $K = 3$ and $R = 0.001$, and with 13B models $K = 3$, $R = 0.023$ are used. For VTW, we use $K = 4$ for LLaVA 1.5 models and $K = 3$ for LLaVA 1.6 model. For M^3, S is set to 56, and for FitPrune pruning ratio is set to 90%.

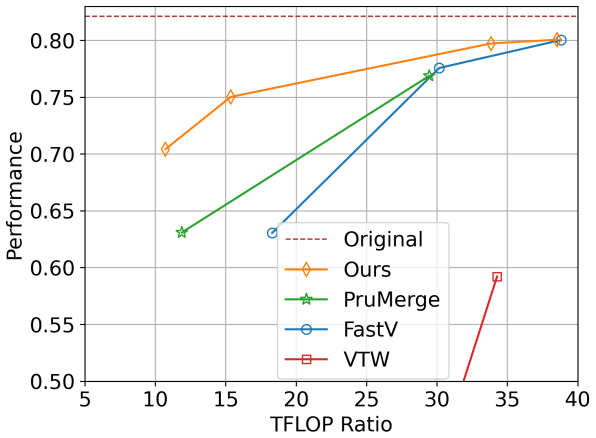


Figure 2. Comparison of different visual token pruning methods across various pruning ratios for LLaVA 1.5-13B. The y-axis is the performance averaged on COCO (CIDEr), OKVQA (Acc), POPE (F1), and MMBench (Acc). The x-axis is the TFLOP ratio of the model after token pruning compared to the original model before pruning.

Prompt: Provide a one-sentence caption for the provided image.

GT caption: A bathroom with a bath tub near windows.

Original Model: A bathroom with a large bathtub and a window.

FastV: A person is standing in front of a white wall.

VTW: A person is standing in front of a painting of a forest.

Ours: A bathroom with a large window and a bathtub.

(a)

Prompt: Provide a one-sentence caption for the provided image.

GT caption: A giraffe and a zebra are on a grassy field by the water.

Original Model: A giraffe and a zebra are standing in a grassy field.

FastV: A person is standing in front of a computer screen.

VTW: A person is standing in front of a wall with a painting \ of a forest on it.

Ours: A giraffe and a zebra are standing in a grassy field.

(b)

Prompt: Provide a one-sentence caption for the provided image.

GT caption: A car in front of a train on train tracks.

Original Model: A silver car is parked in front of a silver train.

FastV: A person is standing in front of a computer screen.

VTW: A person is standing in front of a computer monitor.

Ours: A silver car is parked next to a train.

(c)

Figure 3. Visual examples for image captioning task, comparing the model outputs using the baselines and the proposed method. Colors in text are used for better readability.

Prompt: When the provided information is insufficient, \
respond with 'Unanswerable'. Answer the question using \
a single word or phrase.

Question: What kind of skiing is this person engaged in?

GT answer: Cross country

Original Model: Cross country

FastV: Downhill

VTW: Downhill

Ours: Cross country

(a)

Prompt: When the provided information is insufficient, \
respond with 'Unanswerable'. Answer the question using \
a single word or phrase.

Question: What sates are these grown in?

GT answer: Florida, California (either one is correct)

Original Model: Florida

FastV: Florida

VTW: Unanswerable

Ours: Florida

(b)

Prompt: When the provided information is insufficient, \
respond with 'Unanswerable'. Answer the question using \
a single word or phrase.

Question: What character does this clock look like?

GT answer: Mickey mouse, WALL-E (either one is correct)

Original Model: Mickey mouse

FastV: Clock

VTW: Clock

Ours: Mickey mouse

(c)

Figure 4. Visual examples for visual question answering task, comparing the model outputs using baselines and the proposed methods. Colors in text are used for better readability.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8948–8957, 2019. 2

[2] Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. *Proceedings of the International Conference on Learning Representation*, 2025. 1, 2

[3] 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 *European Conference on Computer Vision (ECCV)*, 2024. 1, 2

[4] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 2

[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[6] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 1, 2

[7] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 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. 1, 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. 2

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 2

[11] Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. *arXiv preprint arXiv:2405.05803*, 2024. 1

[12] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer, 2025. 2

[13] Pan Lu, Swaroop Mishra, Tony 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, 2022. 2

[14] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024. 2

[15] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding, 2023. 2

[16] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 4

[17] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123 (1):74–93, 2017. 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. 1, 2

[19] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 1, 2

[20] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786, 2021. 2

[21] Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. *arXiv preprint arXiv:2409.10197*, 2024. 1, 2

[22] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134, 2019. 2

[23] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024. 2

[24] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024. 1