Frequency-Modulated Visual Restoration for Matryoshka Large Multimodal Models

Supplementary Material

This supplementary material contains several sections that provide additional details related to our work. Specifically, it will cover the following topics:

- In Appendix A, we provide details of the experimental setup, including model architectures and evaluation benchmarks.
- In Appendix B, we provide addition experiments, including FMVR for advanced open-source LMM, FMVR for larger language model, Ablation study on other Matryoshka visual token sampling methods, and case results.

A. Details of Experimental Setup

A.1. Model Architectures

LLaVA-1.5 [18]. As a representative line of open-source multimodal large language models, the LLaVA family is widely adopted for its efficiency and strong capability. It mainly has three components, where CLIP [24] is used to extract visual representations, Vicuna [9] serves as the language backbone, and a lightweight projection module maps visual features into the LLM's embedding space. With visual instruction tuning, this basic setup enables the model to perform a wide range of image–text tasks. LLaVA-1.5 significantly enhance multimodal reasoning ability. At the common resolution of 336×336, the model outputs 576 visual tokens for each image.

LLaVA-Next [19]. LLaVA-NeXT (also referred to as LLaVA-1.6) pushes the LLaVA architecture further by introducing a dynamic-resolution strategy aimed at strengthening the model's visual understanding. Instead of relying on a fixed input size, it adjusts the aspect ratio according to the native dimensions of the image and allows the effective resolution to scale by up to four times. Importantly, the visual encoder itself remains unchanged. The high-resolution input is partitioned into multiple tiles, each matching the original input size, and every tile is processed independently before their visual features are concatenated and passed to the language model. This tiling-based high-resolution pipeline substantially boosts performance on tasks requiring fine-grained perception, such as OCR, detailed reasoning, and factual recognition. For consistency in our evaluations, we standardize the resolution to 672×672—four times the original—yielding a total of 2,880 visual tokens.

Qwen2.5-VL [3]. Qwen2.5-VL represents a member of Qwen-VL family. Its visual backbone has been thoroughly redesigned, adopting a revised Vision Transformer that incorporates windowed attention, SwiGLU activations, and RMSNorm, following the architectural paradigm of the Qwen2.5 language model. Beyond static image understanding, it embraces a dynamic visual processing pipeline, supporting flexible input resolutions and adaptive frame rate handling. Thanks to these advantages, Qwen2.5-VL achieves comparable performance in detailed visual perception tasks such as detection, OCR, layout and document parsing, as well as structured information extraction.

A.2. Evaluation Benchmarks

A.2.1. General Image Benchmarks

VQAv2 [10]. It is the version of the VQA benchmark [2], constructed to assess a model's integrated understanding of visual content, natural language, and commonsense reasoning. The dataset contains 265,016 images drawn from COCO [17] as well as synthetic abstract scenes, with each image paired with an average of 5.4 questions. Every question is further provided with 10 reference answers and 3 additional plausible distractor responses. Following common practice, we adopt the test-dev split for evaluation.

GQA [13]. It is a large-scale VQA dataset constructed from real-world images sourced from the Visual Genome corpus [11], aimed at evaluating the model's capacity for compositional reasoning and fine-grained visual comprehension. It contains more than 22 million question–answer pairs, and every image is accompanied by a richly annotated scene graph that captures object categories, attributes, and inter-object relationships. For our experiments, we report results on the test-dev balanced split.

VizWiz [11]. This benchmark targets visual question answering in a real-world accessibility scenarios: blind or low-vision users take photos in everyday environments and verbally pose questions about them. Each visual question is accompanied by ten answers, offering diverse perspectives on noisy and incomplete visual information. The dataset defines two core tasks—providing answers to the visual questions and detecting when a question cannot be answered from the image alone—thereby emphasizing real-world challenges such as low-quality imagery, occlusion, and ambiguous scene content. We evaluate our model on the official test split.

ScienceQA [21]. This dataset is a large-scale multimodal multiple-choice QA benchmark that covers a broad spectrum of scientific disciplines. It comprises 21,208 questions drawn from natural sciences, language sciences, and social sciences, which are further organized into 26 topics, 127 fine-grained categories, and 379 skill types. The questions vary in modality: 48.7% provide visual content, 48.2% include textual context, and 30.8% offer both forms of information. For evaluation, we follow prior works and adopt the test split that contains image-based questions (ScienceQA-IMG).

POPE [16]. This benchmark targets the evaluation of object hallucination in large vision–language models. It is constructed using images from COCO [17], enabling the measurement of hallucinated predictions. The performance is summarized using precision, recall, and F1 scores. Following prior works, we report results on the test split.

MME [6]. It evaluates both the perceptual and cognitive abilities of multimodal large language models across 14 subtasks. The perception tasks cover coarse- and fine-grained recognition as well as OCR, ranging from basic object presence, count, and attributes to identifying specific scenes, landmarks, and artworks. The cognition tasks assess commonsense reasoning, numerical calculation, translation, and code understanding.

MMBench [20]. This benchmark provides a broad evaluation of vision–language abilities across diverse tasks. It offers a larger varied set of questions and skills than prior benchmarks. MMBench also proposes a CircularEval protocol, which uses ChatGPT to convert open-ended outputs into structured choices for more stable and consistent scoring.

MM-Vet [32]. It examines the integration of multiple multimodal abilities. It covers six core capabilities—recognition, OCR, knowledge, language generation, spatial reasoning, and mathematics—through 218 challenging examples. Evaluation is conducted using a ChatGPT-based assessor, which provides unified metrics for answers of different formats.

A.2.2. Text-oriented benchmarks

TextVQA [26]. This benchmark examines the model's capacity to recognize and reason over textual content embedded in images. The visual data, largely drawn from Open Images v3 [15], contains real-world settings such as signs, advertisements, and product labels, all featuring substantial scene text. As a result, the benchmark provides an assessment of OCR integration and text-sensitive visual reasoning. We evaluate on the validation split.

ChartQA [23]. It is a large benchmark for chart-based question answering, emphasizing visual understanding and logical or complex reasoning. It contains 9.6K human-written questions and 23.1K questions generated from chart summaries. Unlike earlier template-driven datasets, ChartQA requires multi-step reasoning over both chart visuals and their underlying data. We use the test split for evaluation.

Table 1. Performance comparison of different pruning methods on Qwen2.5-VL-7B.

Methods	#Vision Tokens	TextVQA	ChartQA	AI2D	MMB^{EN}	Avg. (%)
Qwen2.5-VL	1296	84.8	86.1	80.4	82.8	83.5
FastV	128	73.8	52.2	71.4	72.9	67.6
DivPrune	128	67.0	50.4	72.1	77.8	66.8
CDPruner	128	77.8	59.2	74.0	76.2	71.8
FMVR (Ours)	81	76.3	62.2	77.5	79.6	**73.9**

AI2D [14]. AI2D is a diagram-based question answering benchmark containing over 5,000 grade-school science diagrams, annotated with more than 150,000 structured labels and syntactic parses. It also provides over 15,000 multiple-choice questions paired with the diagrams, supporting research on visual reasoning and scientific diagram understanding. We use the masked test split for evaluation.

A.2.3. Video benchmarks

MSVD [7]. MSVD is a video question answering benchmark, containing roughly 2k short video clips and about 50k question–answer pairs. It evaluates a model's ability to understand visual content over time, including actions, events, and temporal relationships, as well as its capacity to ground and answer natural language queries about the video.

MSRVTT [30]. It is a large-scale video captioning and video understanding dataset consisting of 10k video clips collected from real-world web videos. Each video is paired with 20 captions, covering a broad range of everyday events, activities, and scenes. The dataset is widely used for video-language tasks such as video captioning, video retrieval, and video question answering, serving as a standard benchmark for evaluating multimodal temporal understanding.

ActivityNet [4]. It is a large-scale benchmark for human activity understanding in untrimmed videos. It contains around 20k videos spanning 200 activity categories, covering a wide range of daily human actions. Designed for tasks such as action recognition, temporal action localization, and temporal segmentation, ActivityNet provides diverse and long-duration videos that evaluate both visual perception and temporal reasoning capabilities of multimodal models.

Video-based Generative Performance Benchmark [22]. It contains five dimensions: correctness, detail orientation, contextual understanding, temporal understanding, and consistency. The evaluation pipelines for both the open-ended question-answering and the generative performance benchmarks adhere to Video ChatGPT [22].

Table 2. Performance comparison on LLaVA-1.5-13B across 10 image-based benchmarks. '#VisionTokens' is the number of vision tokens. Δ rows show the change vs. LLaVA-v1.5-13B.

Methods	#Vision Tokens	VQAv2	GQA	VisWiz	SQA^{IMG}	VQA^{Text}	POPE	MME	MMB^{EN}	MMB^{CN}	MMVet	Avg. (%)
LLaVA-1.5-13B [18]	576	80.0	63.3	53.6	72.8	61.2	86.0	1531.2	68.5	63.5	36.2	66.2
FastV [8]	128	75.3	58.3	54.6	74.2	58.6	75.5	1460.6	66.1	62.3	32.8	63.1
PyramidDrop [29]	128	78.2	61.0	53.8	73.3	60.2	83.6	1489.5	67.5	62.8	32.1	64.7
SparseVLM [35]	128	77.6	59.6	51.4	74.3	59.3	85.0	1487.9	68.4	62.6	35.2	64.8
Prumerge+ [25]	128	76.2	58.3	52.8	73.3	56.1	82.7	1445.9	66.3	61.2	33.6	63.3
TRIM [27]	128	76.4	59.4	49.7	72.4	55.0	86.8	1426.9	67.1	58.4	35.1	63.2
VisionZip [31]	128	76.8	57.9	52.3	73.8	58.9	82.7	1449.2	67.4	62.5	36.0	64.1
DART [28]	128	75.7	57.7	53.0	74.2	58.7	80.4	1395.0	65.4	62.2	34.8	63.2
DivPrune [1]	128	77.1	59.2	53.5	72.8	58.0	86.8	1457.7	66.3	60.7	34.4	64.2
CDPruner [34]	128	77.7	59.7	52.9	73.2	58.4	87.3	1478.0	67.5	61.5	36.2	64.8
VisPruner [33]	128	76.1	58.8	53.4	72.8	57.9	86.0	1452.2	66.4	61.3	36.1	64.1
MQT-LLaVA [12]	144	77.9	59.4	54.1	73.6	60.2	86.1	1471.3	66.5	61.0	31.4	64.4
M3 [5]	144	79.2	60.6	53.5	72.9	60.4	87.7	1478.3	67.2	59.8	32.3	64.8
Ours												
	1	72.2	57.6	50.3	69.4	53.9	82.5	1384.7	62.2	55.8	27.8	60.1
	9	76.3	60.2	51.8	70.9	58.4	84.7	1470.5	65.1	60.4	30.7	63.2
FMVR-LLaVA	36	78.4	61.2	53.4	71.5	60.5	86.6	1477.8	66.3	62.5	32.7	64.7
	144	79.7	63.6	56.7	72.0	61.3	87.2	1489.4	68.3	63.8	34.2	66.1
	576	80.7	64.2	57.2	72.4	60.7	88.1	1528.9	69.5	65.1	35.4	**67.0**
Δ vs. LLaVA-v1.5-13B	144	-0.3	+0.3	+3.1	-0.8	+0.1	+1.2	-41.8	-0.2	+0.3	-2.0	-0.1

Table 3. Ablation study on Matryoshka visual token sampling, including average pooling, sequential sampling, spatial sampling, and max pooling.

Num of Vis Tokens	TextVQA				MMBench			
	Avg Pooling	Sequential	Spatial	Max Pooling	Avg Pooling	Sequential	Spatial	Max Pooling
1	49.2	46.3	45.9	47.1	60.7	57.9	58.6	59.8
9	50.8	46.7	46.4	48.2	64.2	60.4	61.5	63.9
36	55.3	51.2	49.0	52.8	65.2	61.8	63.1	64.5
144	55.5	53.7	50.6	53.3	65.8	64.2	63.7	64.9
576	57.8	55.6	52.6	54.7	65.9	64.7	64.0	65.3

B. Additional Experimental Results

B.1. FMVR for advanced open-source LMM

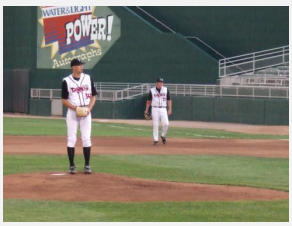
In addition to LLaVA, we further apply FMVR to one of the most advanced open-source LMM, i.e., Qwen2.5-VL. The input image is resized to 1008×1008 and the visual encoder produces a total of 1,296 tokens. We directly insert the trained FMVR into the back of the Qwen2.5-VL's visual encoder. We compare our FMVR with FastV DivPrune, and CDPruner. As shown in Table 1, FMVR consistently achieves the best results across on four benchmarks. FMVR with 81 visual tokens outperforms the next-best method, CDPruner with 128 visual tokens, by 2.1% in average. It demonstrates the strong generalizability of FMVR on advanced LLM architectures.

B.2. FMVR for larger language model

To further examine whether our approach scales to stronger language backbones, we conduct experiments on LLaVA-1.5-13B. As shown in Table 2, Increasing the size of LLM yields substantial performance gains. Across all token pruning methods with 128 visual tokens except for SparseVLM, CDPruner, and M3, FMVR with only 36 visual tokens consistently achieves the highest performance 64.7%. It is worth noting that our FMVR with 144 visual tokens achieved comparable performance to LLaVA-1.5-13B with 576 visual tokens, i.e., 66.1% vs. 66.2%, demonstrating its effectiveness on larger language models.

Table 4. Baseline results of token numbers on LLaVA-1.5-7B.

Tokens	Method	VQAv2	GQA	VisWiz	SQA^{IMG}	VQA^{Text}	POPE	MME	MMB^{EN}	MMB^{CN}	MMVet
576	baseline	78.5	62.0	50.0	66.8	**58.2**	85.9	**1510.7**	64.3	**58.3**	30.5
	Ours	**79.2**	**63.0**	**56.5**	**68.9**	57.8	**87.5**	1510.1	**65.9**	58.0	**34.3**
144	baseline	77.8	61.1	48.3	67.5	**57.3**	85.0	**1480.4**	65.1	57.1	30.2
	Ours	**78.6**	**62.3**	**55.1**	**69.7**	55.5	**86.4**	1473.9	**65.8**	**57.6**	**33.4**
36	baseline	76.1	58.4	46.8	**67.0**	**56.7**	85.2	1448.3	64.5	57.7	29.4
	Ours	**76.5**	**60.9**	**52.9**	69.5	55.3	**85.9**	**1452.5**	**65.2**	**58.3**	**32.2**
9	baseline	73.4	56.3	46.0	67.3	**52.2**	83.6	**1422.9**	63.7	56.8	28.5
	Ours	**74.5**	**59.1**	**50.7**	**69.9**	50.8	**84.1**	1415.0	**64.2**	**57.5**	**29.0**
1	baseline	67.4	54.2	45.5	66.4	**49.6**	79.5	**1291.0**	58.2	**53.9**	26.2
	Ours	**68.3**	**55.2**	**49.7**	**68.6**	49.2	**81.1**	1284.8	**60.7**	53.4	**26.4**



Instruction:
what does the sign say in the background?

Ours (144 tokens): water & light power autographs

M3 (144 tokens): Water's Light! Auto Power

LLaVA-v1.5 (576 tokens): Water & Light Power

Figure 1. Example demonstrating FMVR's image understanding capability on more challenging OCR task. Response marked in red indicates errors.

B.3. Ablation study on other Matryoshka visual token sampling methods

We evaluate four strategies for selecting visual tokens in Matryoshka visual token construction, including average pooling, spatial sampling, sequential sampling, and max pooling. As illustrated in Table 3, average pooling consistently outperforms the other three approaches because it preserves the complete local semantic information by aggregating all tokens within each spatial region. In contrast, spatial sampling and sequential sampling select only one token from each region or sequence, which discards important information and disrupts spatial consistency. Max pooling keeps only the strongest activation, ignoring other semantic information. As a result, average pooling provides the most informative representation, leading to superior performance.

B.4. Additional case results

Fig. 1 presents an example that highlights FMVR's strength on a challenging OCR task. The text in the image varies in size and is somewhat obscured, which typically degrades recognition performance in LLMs. Despite this difficulty, our FMVR correctly read all the text using only 144 visual tokens. In contrast, M3 misidentify key words, such as mistakenly identifying 'autographs' as'auto power' and introduce an incorrect "s'. Furthermore, LLaVA-v1.5 fails to recognize 'autographs'. These results demonstrate that FMVR possesses more detailed visual perception capabilities.

B.5. Baseline results of token numbers

Table 4 reports the baseline and our method under different token budgets on LLaVA-1.5-7B. As token numbers decrease from 576 to 1, performance gradually drops. However, our method consistently outperforms the baseline on most benchmarks, demonstrating stronger robustness and effectiveness under aggressive token reduction.

References

[1] Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9392–9401, 2025. 3

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 1

[4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 2

[5] Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. Matryoshka multimodal models. *arXiv preprint arXiv:2405.17430*, 2024. 3

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

[7] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 2

[8] 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*, pages 19–35. Springer, 2024. 3

[9] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023. 1

[10] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1

[11] 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

[12] Wenbo Hu, Zi-Yi Dou, Liunian Li, Amita Kamath, Nanyun Peng, and Kai-Wei Chang. Matryoshka query transformer for large vision-language models. *Advances in Neural Information Processing Systems*, 37:50168–50188, 2024. 3

[13] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 1

[14] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer, 2016. 2

[15] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2(3):18, 2017. 2

[16] 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. 2

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 2

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 3

[19] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, january 2024. *URL https://llava-vl. github. io/blog/2024-01-30-llava-next*, 2024. 1

[20] 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, 2024. 2

[21] 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. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 2

[22] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad 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 (Volume 1: Long Papers)*, pages 12585–12602, 2024. 2

[23] Ahmed Masry, Xuan Long Do, 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. 2

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[25] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22857–22867, 2025. 3

[26] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 2

[27] Dingjie Song, Wenjun Wang, Shunian Chen, Xidong Wang, Michael X Guan, and Benyou Wang. Less is more: A simple yet effective token reduction method for efficient multimodal llms. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7614–7623, 2025. 3

[28] Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. Stop looking for important tokens in multimodal language models: Duplication matters more. *arXiv preprint arXiv:2502.11494*, 2025. 3

[29] Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*, 2024. 3

[30] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 2

[31] Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19792–19802, 2025. 3

[32] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: evaluating large multimodal models for integrated capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, pages 57730–57754, 2024. 2

[33] Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20857–20867, 2025. 3

[34] Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. *arXiv preprint arXiv:2506.10967*, 2025. 3

[35] Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024. 3