VFlowOpt: A Token Pruning Framework for LMMs with Visual Information Flow-Guided Optimization

Supplementary Material

A. Overview of Baselines

- FastV[3] is a plug-and-play method that optimizes inference efficiency in LMMs by dynamically pruning visual tokens after the second layer, significantly reducing computational costs while maintaining performance. It identifies that image tokens receive drastically lower attention in LLM and strategically removes less impactful tokens.
- FitPrune [16] is a training-free method for pruning visual tokens in multimodal LMMs, based on quickly estimating optimal pruning schemes through attention distribution fitting. It statistically determines which tokens can be discarded by minimizing divergence between attention distributions before and after pruning, using only a small batch of inference data. This approach rapidly produces a pruning recipe tailored to a given computation budget, significantly reducing computational complexity while preserving model performance.
- Pdrop [14] accelerates large vision-language models by progressively removing redundant visual tokens in deeper layers based on token similarity. It partitions models into multiple stages, maintaining all tokens initially to preserve critical visual information, then gradually pruning tokens as layers deepen. This approach effectively reduces computational costs without compromising performance during both training and inference.
- Sparsevlm [17] introduces a training-free, text-guided visual token sparsification method for LMMs, significantly reducing computational overhead by adaptively selecting important visual tokens based on relevant text prompts. It employs an adaptive pruning strategy at each layer and recycles pruned visual tokens into compact representations to minimize information loss.
- Visionzip [15] is a simple yet effective method that reduces visual token redundancy in LMMs by selecting only the most informative tokens, significantly improving efficiency while maintaining performance. It employs a text-agnostic approach that merges and compresses redundant tokens, reducing computational costs and enhancing inference speed without requiring additional training.

B. Overview of Benchmarks

- MME [4] offers a robust benchmark for evaluating LVLMs across multimodal tasks. It assesses models on two major fronts: perception and cognition, using 14 well-structured subtasks that challenge their interpretive and analytical abilities.
- MMBench [10] takes a two-pronged approach by introducing an extensive dataset that broadens the scope of evaluation questions and a novel CircularEval strategy that utilizes ChatGPT to convert free-form responses into structured answer choices.
- ScienceQA [11] focuses on evaluating multi-hop reasoning and interpretability within scientific domains. It features a large dataset of approximately 21K multiple-choice questions across a variety of science topics, accompanied by detailed annotations and explanations.
- VizWiz [7] stands out in the VQA field by using a dataset of over 31,000 visual questions that come from a real-world setting, featuring images taken by visually impaired individuals and their associated spoken queries, along with crowdsourced answers.
- GQA [1] is built for complex visual reasoning tasks, containing 22 million questions generated from scene graph-based structures. It incorporates innovative evaluation metrics focused on consistency, grounding, and plausibility, pushing the boundaries of vision-language evaluation.
- POPE [9] introduces a methodology to evaluate object hallucination in LVLMs, transforming the task into a binary classification problem. By using simple Yes-or-No prompts, POPE highlights model tendencies towards hallucination through various object sampling strategies.
- VQA [6] collects complementary images such that every question in the balanced dataset is associated with a pair of similar images that result in two different answers to the question.
- ChartQA [12] is a large-scale benchmark designed for question answering on charts, focusing on both visual and logical reasoning with 9.6K human-written and 23.1K automatically generated questions.
- DocVQA [13] is a large-scale dataset designed for Visual Question Answering (VQA) on document images, containing 50,000 questions over 12,000+ real-world documents. Unlike previous datasets, it requires models to understand both textual content and visual layout, including tables, forms, and complex structures.
- MMstar [2] is a new benchmark designed to address issues in evaluating Large Vision-Language Models (LVLMs), specifically unnecessary visual content and unintentional data leakage, which can mislead performance assessments. It includes 1,500 carefully selected vision-dependent samples, ensuring accurate evaluation of LVLMs' true multi-modal reasoning abilities. MMStar

Methods	Token Reduction	FLOPs ↓ (T)	Δ	Latency ↓ (ms)	Δ	KV Cache ↓ (MB)	Δ	Performance ↑	Δ
LLaVA-OneVision-7B	-	71.4	-	1040.1	-	1786.4	-	1581	-
+ VFlowOpt	50%	37.2	**-48.0%**	584.2	**-43.8%**	902.8	**-49.5%**	1591	+0.6%
+ FastV	50%	38.1	**-46.6%**	615.1	**-41.9%**	902.8	**-49.5%**	1549	-2.0%
+ VisionZip	50%	37.7	**-47.2%**	580.7	**-44,2%**	902.8	**-49.5%**	1587	+0.1%

Table 1. Efficiency analysis of LLaVA-OneVision-7B with VFlowOpt, FastV, and VisionZip. The detailed metric includes computation (FLOPs), latency, and KV-Cache memory. (Δ) denotes the reduction ratio.

	MMStar	MME	MMB	SQA	POPE	GQA	DocVQA	VQA^{Text}
VisionZip	54.6	1562	78.9	90.4	88.8	61.0	79.6	70.0
Ours (Random)	57.8	1570	79.9	92.3	89.1	61.2	82.3	72.5
Ours ($_{MathV360K-GEOS}$)	57.8	1566	79.8	92.0	89.1	61.0	82.1	72.8

Table 2. Impact of optimization data selection

introduces new metrics—Multi-Modal Gain (MG) and Multi-Modal Leakage (ML)—to measure actual improvements from multi-modal training, with evaluations showing GPT-4V leading in both accuracy and multi-modal efficiency.

- SeedBench [8] is a large-scale benchmark designed to evaluate the generative comprehension capabilities of Multimodal Large Language Models (MLLMs), featuring 19K human-annotated multiple-choice questions across 12 evaluation dimensions for both images and videos.
- VideoMME [5] is the first comprehensive benchmark designed to evaluate Multi-Modal Large Language Models (MLLMs) in video analysis, covering 900 manually annotated videos across six diverse domains and 30 subcategories. It introduces a full-spectrum evaluation with multi-modal inputs, including subtitles and audio, and assesses models across various temporal contexts, from short clips to hour-long videos.

C. Efficiency Analysis about Baselines

We evaluate VFlowOpt, the well-performing baseline FastV, and VisionZip on efficiency metrics under the condition of retaining 50% of the tokens. With the same token retention rate, all methods showed identical KV-Cache memory usage, while FLOPs and latency exhibited slight differences, as shown in Tab. 1.

D. More ablation studies

D.1. Choice of the optimization target

We are inspired by previous interpretability studies (Main Paper L273–L281) and consider the last token as the most representative one of such interactions. Results (shown in Tab. 3) show that optimizing for the last token yields the best performance. We will add this in the revised paper.

	MMStar	MME	MMB	SQA	POPE	GQA
Last Token	**57.8**	**1570**	**79.9**	**92.3**	**89.1**	**61.2**
Mean Pooling	56.1	1549	77.5	92.1	88.5	60.6
First Token	54.2	1530	77.7	89.5	85.4	60.4
Top-3 Tokens	56.8	1544	78.6	92.3	88.3	61.1

Table 3. Analysis of choice of the optimization target

	DocVQA	VQA^{Text}	POPE
VFlowOpt	82.3	72.5	89.1
w/o Importance Calibration	80.3	71.4	88.6
w/o Token Merging	82.0	72.4	86.8
w/o Progressive Pruning	81.9	71.6	88.2

Table 4. Ablation studies on more benchmarks

D.2. Impact of optimization data selection

The result of our optimization is independent of data selection because the visual information flow being optimized is task-agnostic and model-specific. In our experiments, repeated random sampling yields nearly identical results. To further validate this, we optimize using 30 samples from the task-specific split (MathV360K-Geometry3K) of the LLaVA-OV training data. The model consistently achieves strong results across various tasks, regardless of data selection (shown in Tab. 2).

D.3. Ablation studies on more benchmarks

Additional results in Tab. 4 show that Token Merging is crucial for preserving coarse-grained semantics, while Importance Calibration and Progressive Pruning help maintain fine-grained visual perception.

References

[1] Joshua Ainslie, James Lee-Thorp, Michiel De Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023. 1

[2] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 1

[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*, pages 19–35. Springer, 2024. 1

[4] 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, 2024. 1

[5] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 2

[6] 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

[7] 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

[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] 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. 1

[11] 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. 1

[12] 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. 1

[13] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 1

[14] 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. 1

[15] 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. *arXiv preprint arXiv:2412.04467*, 2024. 1

[16] 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

[17] 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. 1