IF-Prune: Information-Flow Guided Token Pruning for Efficient Vision-Language Models

Supplementary Material

1. Experiments

1.1. IF-Prune on Different Model Architecture

In addition to the experiments on InternVL, we further evaluate the generalizability of IF-Prune by applying it to LLaVA-1.5 [3]. As shown in Table T1 and Table T5, we compare IF-Prune against several strong baselines, including ToMe [1], FastV [2], PyramidDrop [7], SparseVLM [9], HiPrune [4], and VisionZIP [8].

Table T1. Comparison between IF-Prune and recent baselines on LLaVA-1.5-7B. Please refer to the supplementary for more results.

Method	GQA	MMB	POPE	SQA	MME	Average
LLaVA-1.5-7B	61.9	64,7	86.9	69.5	1862	100%
Retain 64 tokens (11.1% token ratio)						
ToMe	48.7	43.7	52.5	50.0	-	69.1%
FastV	46.1	48.0	48.0	51.1	1256	67.6%
PyramidDrop	47.5	58.8	55.9	69.2	1561	83.6%
VisionZip	55.1	60.1	77.0	69.0	1690	91.0%
SparseVLM	53.7	60.1	77.5	69.7	1559	84.8%
HiPrune	53.6	59.5	73.0	68.9	1646	88.6%
SGP	58.3	60.1	86.6	68.9	1610	87.8%
IF-Prune(ours)	**58.4**	**61.4**	**86.7**	**69.7**	**1731**	**93.6%**

While both IF-Prune and VisionZIP involve a learning stage, the nature and cost of training are fundamentally different. VisionZIP requires fine-tuning the answer model on the pruned token sequences, effectively adapting the entire multimodal backbone to a new compressed visual input distribution. This increases computational overhead and limits deployment flexibility. For a fair comparison, we report the official results of untrained VisionZIP.

As shown in Table T5, IF-Prune surpasses all other baselines under the most aggressive pruning setting (retaining only 64 visual tokens), which aligns with our observations on the InternVL experiments. When retaining 128 visual tokens, IF-Prune achieves performance competitive with VisionZIP.

1.2. More Results

Applying token pruning in vision–language models (VLMs) often degrades performance on fine-grained visual understanding tasks, particularly on text-intensive datasets. To evaluate the robustness of our method in such scenarios, we conduct experiments on TextVQA [6] and DocVQA [5]. As shown in Table T2, when retaining only 5%–10% of visual tokens, IF-Prune consistently outperforms SGP while maintaining comparable inference latency.

We report results on InternVL2-8B and 26B in the main paper, and we add InternVL2-2B in Table T3.

Table T2. Performance comparison between SGP and IF-Prune on InternVL2 with comparable inference time.

Method	InternVL2-2B		InternVL2-8B		InternVL2-26B	
	DocVQA	InfoVQA	DocVQA	InfoVQA	DocVQA	InfoVQA
Baseline	84.9	53.0	90.0	66.5	90.5	68.9
SGP	80.3	48.4	82.7	55.8	83.5	57.4
IF-Prune	**81.2**	**49.0**	**82.9**	**56.5**	**84.0**	**59.2**

Table T3. Performance comparison based on InternVL2-2B.

Method	K	GQA	MMStar	MMBench	VQA^{text}	Average
InternVL2-2B	100%	59.9	48.4	72.7	72.5	100%
SGP	5%	57.8	43.8	65.5	**69.3**	93.3%
IF-Prune	5%	**58.1**	**44.3**	**69.8**	68.6	**95.0%**

Table T4. Hyperparameters for training.

LoRA alpha	64
LoRA rank	32
Batch size (SFT)	16
Gradient accumulation (SFT)	11
Learning rate	0.00005
Optimizer	AdamW
Weight decay	0.01
γ (KL warmup)	$0.2 * total\ steps$
τ_{max}	0.5
τ_{min}	0.2

1.3. Qualitative results

In Fig. 1, Fig. 2, and Fig. 3, we provide a comparison of the importance map and the pruning guidance provided by SGP and our proposed IF-Prune. IF-Prune tends to assign a higher importance score to a wider and more relevant visual tokens than SGP. In Fig. 4 and Fig. 5, we present the qualitative results of IF-Prune on the MMStar dataset, demonstrating the generalizability of IF-Prune on real-world images.

1.4. Hyperparameter

We detail the hyperparameters used for training the information bottleneck in Table T4. To be noticed, the large model was not involved during training. The only trainable parts include the LoRA and the information bottleneck module for the pruning guidance generator.

References

[1] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging:

Table T5. Performance of IF-Prune on LLaVA-1.5 7B. All the methods focus solely on the token pruning part. The best results are **bold**, and the second-best results are underlined.

Method	GQA	MMB	MME	POPE	SQA	MMMU	SEED	MMVet	LLaVA-B	Avg.
Upper Bound, 576 Tokens (100%)										
LLaVA-1.5	61.9 100%	64.7 100%	1862 100%	85.9 100%	69.5 100%	36.3 100%	58.6 100%	31.1 100%	66.8 100%	100%
Retain 64 Tokens (↓88.9%)										
FastV	46.1 74.5%	48.0 74.2%	1256 67.5%	48.0 55.9%	51.1 73.5%	34.0 93.7%	51.9 88.6%	25.8 83.0%	46.1 69.0%	75.6%
SparseVLM	52.7 85.1%	56.2 86.9%	1505 80.8%	75.1 87.4%	62.2 89.4%	32.7 90.1%	51.1 87.2%	23.3 74.5%	57.5 86.1%	85.8%
VisionZip	55.1 89.0%	61.0 92.9%	1690 90.8%	77.0 89.6%	69.0 99.3%	36.2 99.7%	52.2 89.1%	31.7 101.9%	62.9 94.2%	<u>94.0%</u>
IF-Prune	58.4 94.3%	61.4 94.9%	1731 93.0%	86.7 100.9%	69.7 100.3%	34.9 96.1%	50.7 86.5%	26.3 84.6%	80.6 120.7%	**96.8%**
Retain 128 Tokens (↓77.8%)										
FastV	49.6 80.1%	56.1 86.7%	1490 80.0%	59.6 69.4%	60.2 86.6%	34.9 96.1%	55.9 95.4%	28.1 90.9%	52.0 77.8%	83.5%
SparseVLM	56.0 90.5%	60.0 92.7%	1696 91.1%	80.5 93.7%	67.1 96.5%	33.8 93.1%	53.4 91.1%	30.0 96.5%	62.7 93.9%	93.4%
VisionZip	57.6 93.1%	62.0 95.8%	1762 94.6%	83.2 96.9%	68.9 99.1%	37.9 104.4%	54.9 93.7%	32.6 104.8%	64.8 97.6%	<u>97.6%</u>
IF-Prune	60.0 96.9%	62.5 96.6%	1735 93.2%	87.3 101.6%	70.0 100.7%	35.4 97.5%	51.8 88.4%	27.4 88.1%	81.3 121.7%	**98.3%**
Retain 192 Tokens (↓66.7%)										
FastV	52.7 85.1%	61.2 94.6%	1612 86.6%	64.8 75.4%	67.3 96.8%	34.3 94.5%	57.1 97.4%	27.7 89.7%	49.4 74.0%	88.2%
SparseVLM	57.6 93.1%	62.5 96.6%	1721 92.4%	83.6 97.3%	69.1 99.4%	33.8 93.1%	55.8 95.2%	31.5 101.3%	66.1 99.0%	96.4%
VisionZip	59.3 95.8%	63.0 97.4%	1783 95.7%	85.3 99.3%	68.9 99.1%	36.6 100.8%	56.4 96.2%	31.7 101.9%	67.7 101.3%	**98.5%**
IF-Prune	60.3 97.4%	64.0 98.9%	1740 93.4%	87.1 101.4%	69.6 100.1%	35.1 96.7%	52.1 88.9%	26.8 86.2%	78.0 116.8%	<u>97.8%</u>

Your vit but faster. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[2] 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*, 2024. 1

[3] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 1

[4] Jizhihui Liu, Feiyi Du, Guangdao Zhu, Niu Lian, Jun Li, and Bin Chen. Hiprune: Training-free visual token pruning via hierarchical attention in vision-language models. *arXiv preprint arXiv:2508.00553*, 2025. 1

[5] Minesh Mathew, Dimosthenis Karatzas, R. Manmatha, and

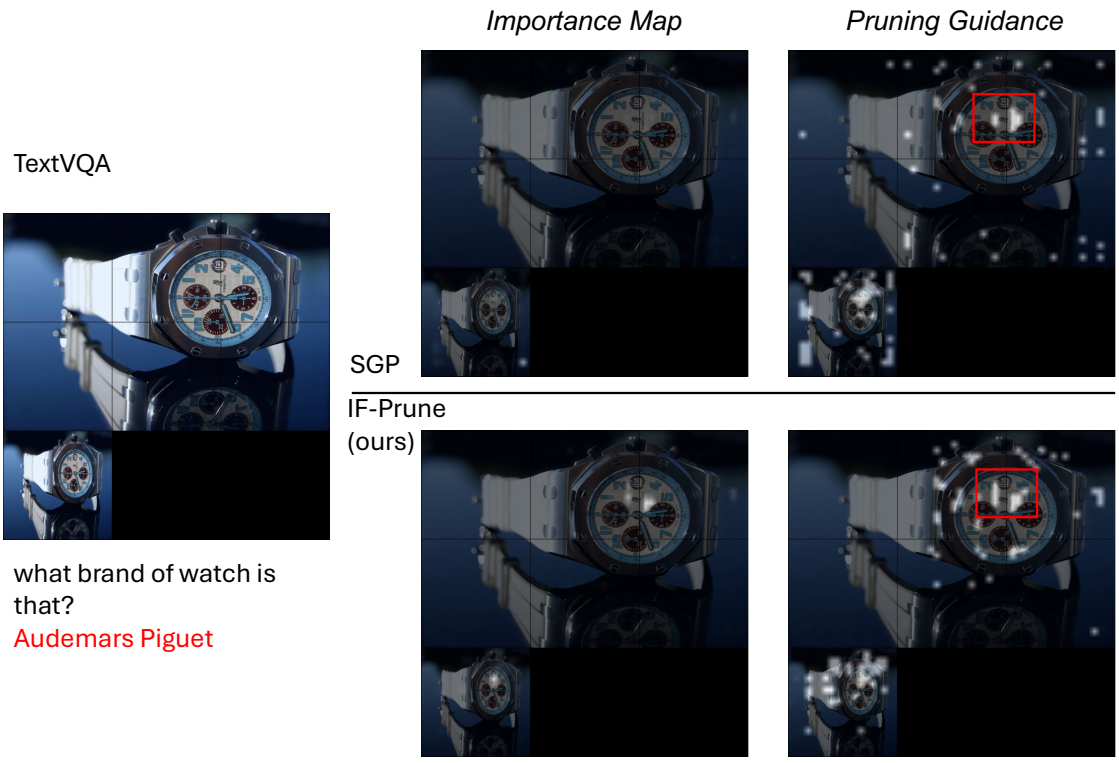

Figure 1. Comparison of the importance map and pruning guidance proposed by SGP and IF-Prune (ours) based on the user input.

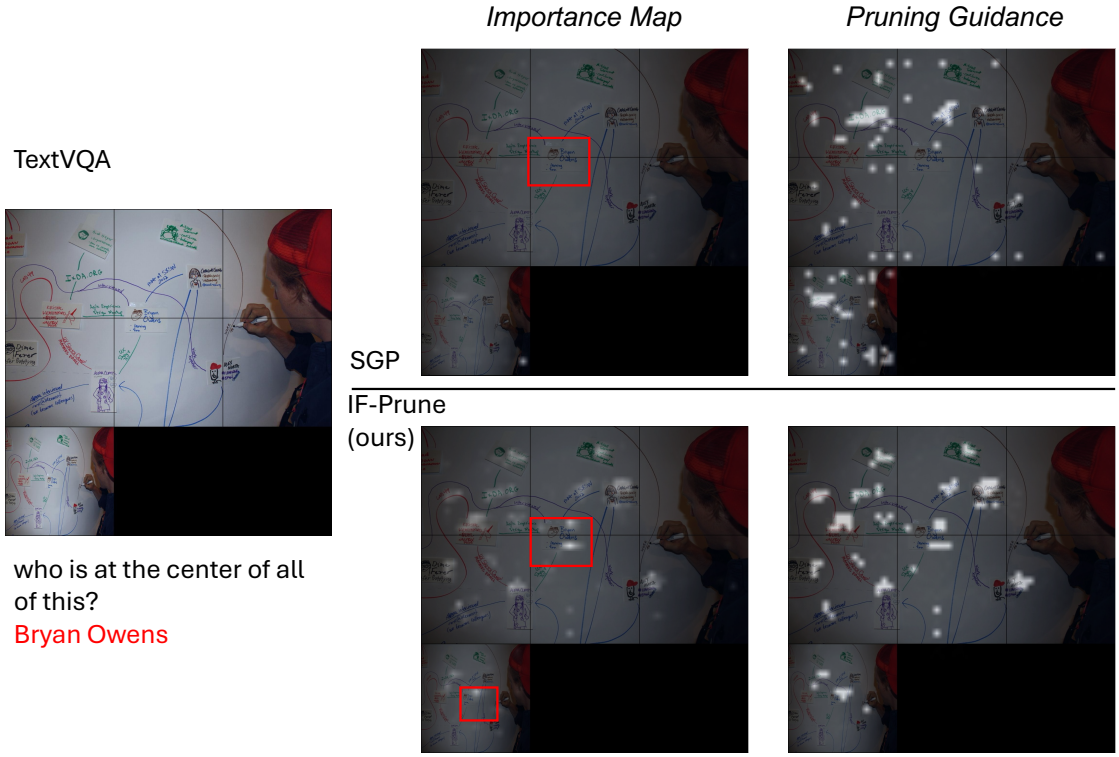

Figure 2. Comparison of the importance map and pruning guidance proposed by SGP and IF-Prune (ours) based on the user input.

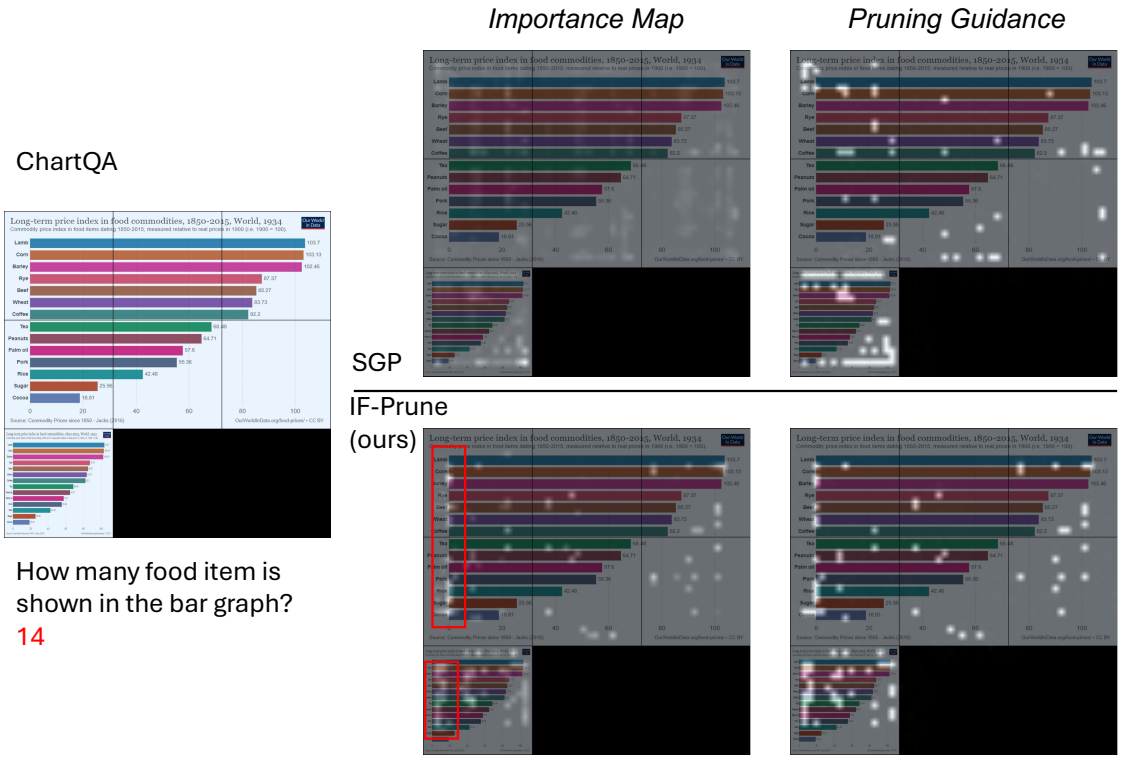

Figure 3. Comparison of the importance map and pruning guidance proposed by SGP and IF-Prune (ours) based on the user input.

Figure 4. Visualization of the importance map and pruning guidance proposed by IF-Prune (ours) based on the user input.

C. V. Jawahar. Docvqa: A dataset for vqa on document images. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2199–2208, 2020. 1

[6] 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*, 2019. 1

[7] Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang,

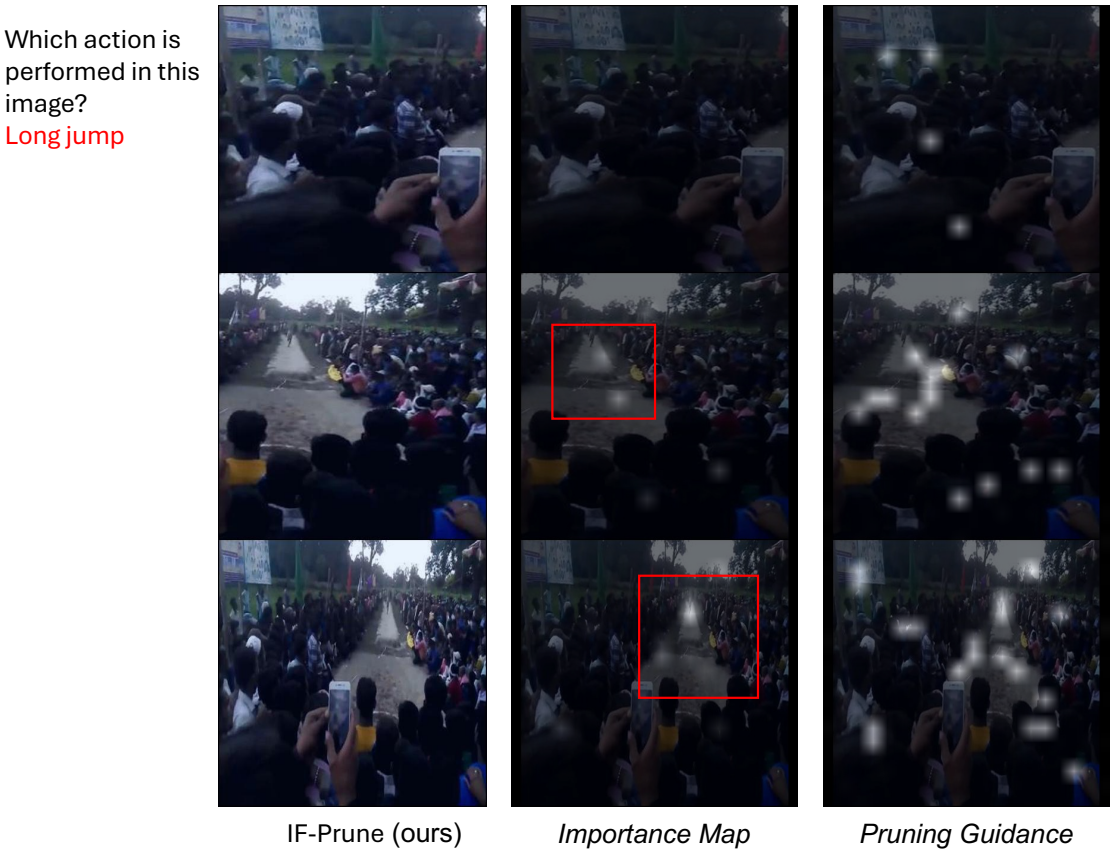

Figure 5. Visualization of the importance map and pruning guidance proposed by IF-Prune (ours) based on the user input.

Feng Wu, et al. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. In *CVPR*, 2025. 1

[8] 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*, 2025. 1

[9] 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. In *ICML*, 2025. 1