QuietPrune: Query-Guided Early Token Pruning for Vision-Language Models

Supplementary Material

1. Pruning configurations for different methods

The original pruning configurations are different across methods. For a fair comparison, we evaluate all methods under the same average visual token pruning rate at the LLM stage. PACT [5] and AIM [8] also incorporate token merging to reduce the number of visual tokens. PACT employs a distance bounded density peaks clustering method for merging the visual tokens after pruning. AIM uses ToMe [3] to merge the visual tokens within the input of LLM. However, token merging methods are time-consuming. For example, in PACT, when applying token pruning to remove 50% of the visual tokens in the Qwen3-VL-4B model, the average prefill latency on OCRBench [6] is 265 ms. In contrast, using token merging to reduce 50% of the tokens results in an average prefill latency of 397 ms, increasing the latency by 49.8%. Moreover, token merging and token pruning are not inherently coupled, so token merging can be combined with any token pruning methods. Therefore, we only compared the token pruning component of PACT and AIM.

For a VLM model, whose LLM component has L transformer layers, the average pruning rate R^* is defined as: $R^* = 1 - \frac{\sum_{i=1}^{L} \hat{T}_i}{\sum_{i=1}^{L} T_i}$, where T_i and $\hat{T}_i$ are the number of visual tokens in layer i with and without pruning, respectively. We compare results in the paper under seven different average pruning rates: 20%, 30%, 40%, 50%, 60%, 70%, and 80%. The specific pruning configurations for each method under different average pruning rates are provided as follows:

QuietPrune (Ours): We apply token pruning at three fixed layers of the ViT, namely at depths $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{3}{4}$ of the total number of blocks. At each of these layers, a fraction R of visual tokens is pruned based on group relevance scores. For different average pruning rates R^*, the specific values of R are as follows:

- $R = 58.5\%$, when $R^* = 20\%$
- $R = 66.9\%$, when $R^* = 30\%$
- $R = 73.7\%$, when $R^* = 40\%$
- $R = 79.4\%$, when $R^* = 50\%$
- $R = 84.3\%$, when $R^* = 60\%$
- $R = 88.8\%$, when $R^* = 70\%$
- $R = 92.8\%$, when $R^* = 80\%$

FastV [4]: FastV prunes the visual tokens in the filtering layer K of the LLM with pruning ratio R. According to the experimental results of FastV, the best configuration is $K = 2$ and $R = 50\%$. To match this, under different average pruning rates, we satisfy one of the optimal conditions:

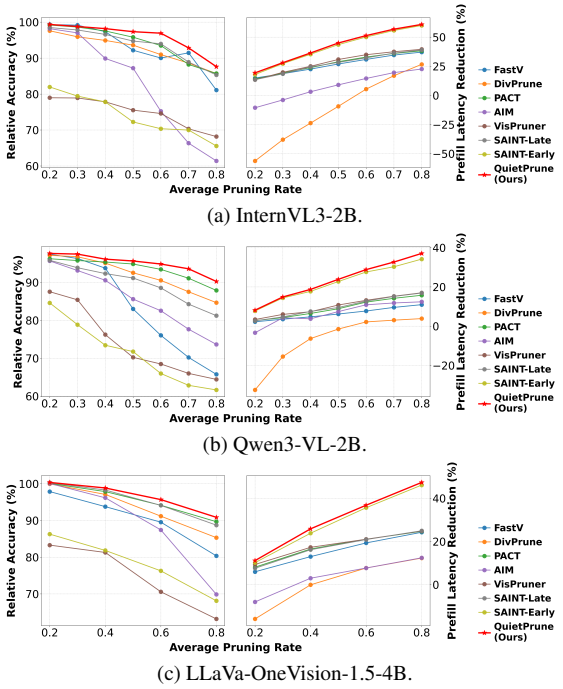
- When $R^* > 50\%$, we fix $R = 50\%$ and adjust K according to the total number of layers in the LLM.
- When $R^* \leq 50\%$, we fix $K = 2$ and adjust R according to the total number of layers in the LLM.

Take InternVL3-8B [9] as an example, whose total number of layers in the LLM is 28. When $R^* = 20\%$, we set $K = 2$ and $R = 86.2\%$. When $R^* = 80\%$, we set $K = 17$ and $R = 50\%$.

DivPrune [1] and VisPruner [7]: In DivPrune and VisPruner, visual tokens are pruned by R before being passed to the first decoder layer in the LLM. Therefore, we directly set $R = R^*$ to reach the target average pruning rate.

PACT [5]: PACT performs token pruning in a reduction layer L with a pruning percentage λ. Following the settings in PACT, we set $L = 4$ for Qwen3-VL series and $L = 7$ for InternVL3 series, and adjust λ according to R^*.

AIM [5]: AIM prunes the visual tokens starting at the L_1 layer of the LLM and totally removes all the visual tokens after layer L_2. According to AIM, the best setting is $L_2 = L_1 + 8$. Following this setting, we set the pruning ratio $R = 12.5\%$ at each pruning layer and adjust L_1 for different average pruning ratios.



(a) InternVL3-2B.

(b) Qwen3-VL-2B.

(c) LLaVa-OneVision-1.5-4B.

Figure 1. Comparison of different methods on the InternVL3-2B and Qwen3-VL-2B. Left is the relative accuracy, and right is the prefill latency reduction of different methods under various pruning rates.

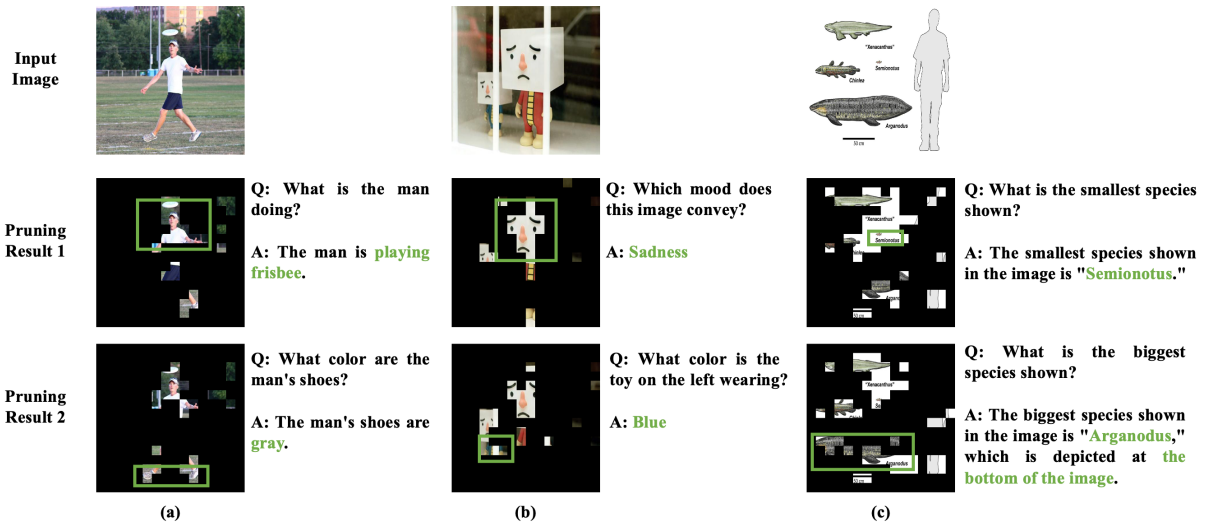


Figure 2. QuietPrune results on InternVL3-8B model. Our method adaptively retains the visual tokens that are relevant to each query.

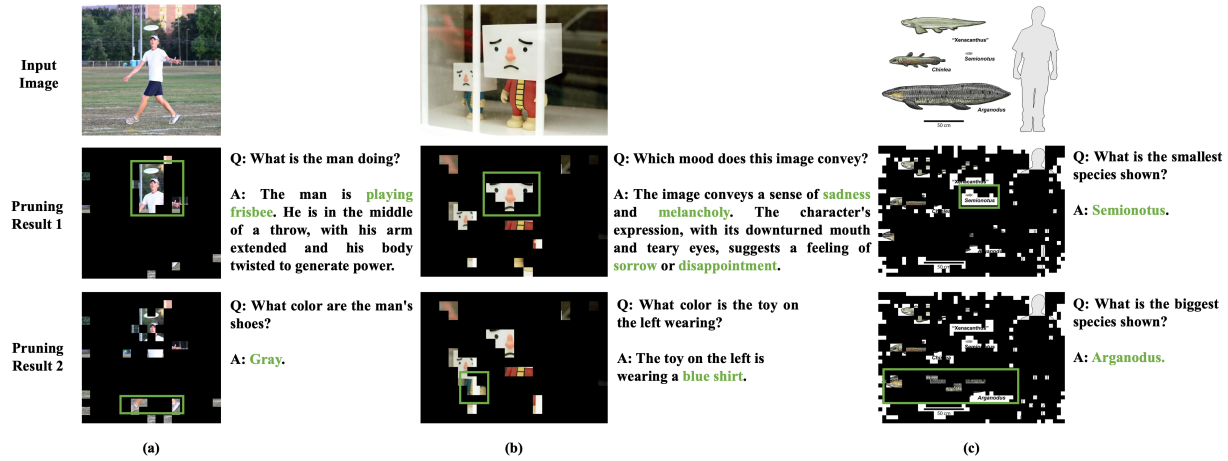


Figure 3. QuietPrune results on Qwen3VL-4B model. Our method adaptively retains the visual tokens that are relevant to each query.

SAINT [5]: For SAINT, we report both its early pruning and late pruning variants. For early pruning, we use the same configuration as our method. For late pruning, the best setting reported in SAINT is to perform progressive pruning at layers $8 - 16$. We follow these settings and control the pruning rate of each layer to reach the average pruning ratio R^*.

2. Additional experiment results

We present the additional experiment results on InternVL3-2B, Qwen3-VL-2B and LLaVa-OneVision-1.5-4B [2] as shown in Fig. 1. Together with the results in Fig. 6 of the main paper, our proposed QuietPrune consistently outperforms existing SOTA methods on both relative accuracy and latency reduction. This result further demonstrates the robustness and generalization capability of our method.

3. More visualization of pruning results

We visualize additional token pruning results of QuietPrune in Fig. 2 and Fig. 3. In these figures, the green bounding boxes highlight the visual tokens retained that are relevant to each query. Pruning result 1 for images (a) and (b) confirms that our query-guided method can still attend to the target region in the query, even when no salient object is present in the query. Pruning results for image (c) demonstrate that our method can capture subtle changes (i.e., "smallest" and "biggest") in the query and correctly adjust the relevance score of tokens. Moreover, a comparison between Fig. 2 and Fig. 3 reveals that, despite architectural differences, various models consistently retain tokens

from highly similar regions when processing the same input query and image. These results further demonstrate the generalization capability and effectiveness of the proposed method.

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. 1

[2] Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025. 2

[3] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022. 1

[4] 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

[5] Mohamed Dhouib, Davide Buscaldi, Sonia Vanier, and Aymen Shabou. Pact: Pruning and clustering-based token reduction for faster visual language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14582–14592, 2025. 1, 2

[6] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024. 1

[7] 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. *arXiv preprint arXiv:2412.01818*, 2024. 1

[8] 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. 1

[9] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 1