Efficient Vision-Language Reasoning via Adaptive Token Pruning

Xue Li
Scholar42, InfiniPouch LLC
xueli@scholar42.com

Xiaonan Song[*]
Scholar42, InfiniPouch LLC
ssong@scholar42.com

Henry Hu
Labelbox, Inc.
hhu@labelbox.com

Abstract

As vision-language models (VLMs) continue to advance toward real-world deployment in domains such as robotics, autonomous systems, and assistive technologies, their computational and memory demands pose a persistent bottleneck. Existing architectures typically process all visual and textual tokens uniformly, regardless of their contribution to the final prediction, leading to inefficiencies and latency that hinder scalability. In this work, we introduce Adaptive Token Pruning (ATP), a dynamic inference mechanism that identifies and retains only the most informative subset of multimodal tokens based on their contextual relevance. ATP operates by analyzing cross-modal attention distributions at each transformer layer, estimating token importance scores derived from both inter- and intra-modal saliency. Tokens deemed redundant are pruned progressively, allowing the model to focus computation on semantically rich regions and phrases while maintaining alignment across modalities.

Unlike static compression or distillation approaches, ATP adapts to each input instance without modifying the backbone architecture. We propose ATP as a lightweight gating module compatible with popular VLM backbones such as BLIP-2, LLaVA, and Flamingo. Preliminary evaluations across VQAv2, GQA, and COCO Captioning indicate that ATP can reduce inference FLOPs by around 40% and achieve roughly 1.5× speedups in end-to-end latency, with negligible loss (<1%) in task accuracy. Moreover, qualitative analyses suggest that ATP preserves visual grounding and contextual reasoning fidelity, indicating that token pruning can also serve as a lens into model interpretability.

Beyond efficiency, we investigate the robustness of ATP-enhanced models under visual corruption and linguistic perturbation scenarios. Our observations suggest that adaptive pruning tends to suppress spurious correlations and hallucinated features, yielding improved stability across noise conditions. These findings suggest that resource-constrained inference and model reliability are not necessarily competing objectives—adaptive mechanisms can improve both simultaneously. Finally, we discuss how ATP can be integrated into deployment pipelines for multimodal edge computing, emphasizing its role as a general design principle for efficient, robust, and real-time VLM reasoning.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh,

[*] Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 1st Workshop on VLM4RWD.

Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. *arXiv e-prints*, page arXiv:2204.14198, April 2022.

[2] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token Merging: Your ViT But Faster. *arXiv e-prints*, page arXiv:2210.09461, October 2022.

[3] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *arXiv e-prints*, page arXiv:2003.10555, March 2020.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv e-prints*, page arXiv:2010.11929, October 2020.

[5] Song Han, Huizi Mao, and William J. Dally. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. *arXiv e-prints*, page arXiv:1510.00149, October 2015.

[6] Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *arXiv e-prints*, page arXiv:1903.12261, March 2019.

[7] Le Hou, Richard Yuanzhe Pang, Tianyi Zhou, Yuexin Wu, Xinying Song, Xiaodan Song, and Denny Zhou. Token Dropping for Efficient BERT Pretraining. *arXiv e-prints*, page arXiv:2203.13240, March 2022.

[8] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv e-prints*, page arXiv:2301.12597, January 2023.

[9] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not All Patches are What You Need: Expediting Vision Transformers via Token Reorganizations. *arXiv e-prints*, page arXiv:2202.07800, February 2022.

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *arXiv e-prints*, page arXiv:2304.08485, April 2023.

[11] Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. DivPrune: Diversity-based Visual Token Pruning for Large Multimodal Models. *arXiv e-prints*, page arXiv:2503.02175, March 2025.

[12] Yizheng Sun, Yanze Xin, Hao Li, Jingyuan Sun, Chenghua Lin, and Riza Batista-Navarro. LVPruning: An Effective yet Simple Language-Guided Vision Token Pruning Approach for Multi-modal Large Language Models. *arXiv e-prints*, page arXiv:2501.13652, January 2025.

[13] Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. SparseVLM: Visual Token Sparsification for Efficient Vision-Language Model Inference. *arXiv e-prints*, page arXiv:2410.04417, October 2024.