# HiPrune: Training-Free Visual Token Pruning via Hierarchical Attention in Vision-Language Models (Student Abstract)

**Jizhihui Liu** [*], **Guangdao Zhu** [*], **Feiyi Du** [*]

Harbin Institute of Technology, Shenzhen-518055
2023112090@stu.hit.edu.cn

## Abstract

Vision-Language Models (VLMs) encode images into lengthy sequences of visual tokens, leading to excessive computational overhead and limited inference efficiency. In this paper, we study the hierarchical attention pattern in vision encoders and propose **HiPrune**, a training-free and model-agnostic token **Prun**ing framework for VLMs. We identify that middle layers in the vision encoder attend to object-centric regions, while deep layers capture global contextual features. Based on this observation, HiPrune selects tokens based on the attention score from the middle and deep layers. Our method requires no retraining and integrates seamlessly with any ViT-based VLM. Experiments demonstrate that HiPrune achieves outstanding pruning performance, maintaining a balance between efficiency and efficacy.

**Code** — https://github.com/Danielement321/HiPrune

## Introduction

Vision-Language-Model (VLM) commonly comprises a vision encoder (Radford et al. 2021), an adaptor, and an LLM (Touvron et al. 2023). The vision encoder is a vision transformer (ViT) (Dosovitskiy et al. 2021) that encodes the image into a sequence of tokens. These visual tokens are overly long and redundant, accounting for the biggest proportion of inputs. In LLaVA-1.5 (Liu et al. 2024), an image is encoded into 576 tokens, much longer than its textual counterparts. For VLMs that incorporate a native dynamic-resolution encoder, one high-resolution webpage snapshot may require more than 10,000 tokens, resulting in a substantial computational cost and GPU memory allocation.

In this paper, we carry out analyses on the focus of each layer in the vision encoder and propose **HiPrune**, a training-free and model-agnostic visual token pruning method for VLM inference acceleration. We point out that the middle layers of vision encoder pay more attention to tokens that are object-related, while the deep layers focus on tokens encoding rich global information. Based on this conclusion, the core idea of HiPrune is to preserve a compact subset of tokens that collectively retain both fine-grained and global visual information, guided by the hierarchical attention within the vision encoder. Experiments demonstrate the versatility of our conclusion on this hierarchical attention pattern and the efficacy of our token pruning method.

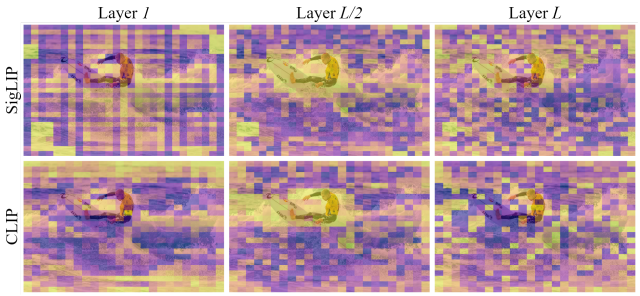


Figure 1: **Attention map for different layers of SigLIP and CLIP.** Patches with higher scores are in yellow. We can see that the middle layer is more object-centric.

## Methodology

### Identify the Hierarchical Attention Pattern in Vision Encoders

In a transformer layer, the attention is computed by

$$\mathbf{A} = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}) \in \mathbb{R}^{H\times N\times N}, \tag{1}$$

where $\mathbf{Q}, \mathbf{K}$ demote query and key, respectively, and $d_k$ is the dimension of key.

For each token, the attention score indicates the proportion it takes up in all the tokens, which is computed by

$$\mathbf{a}^{[l]} = \frac{1}{H}\sum_{h=1}^{H}\sum_{n=1}^{N}\mathbf{A}^{[l]}[h,n,:], \tag{2}$$

$$= (a_1^{[l]}, a_2^{[l]}, \ldots, a_N^{[l]}) \in \mathbb{R}^N. \tag{3}$$

We plot the attention score for CLIP and SigLIP in Fig. 1. In the middle layers, the model concentrates on the main object of the image like the surf-man in Fig. 1. We further compute the IoU between the object segmentation mask and top 10% high-attention tokens in Table 2 on the COCO val2017 dataset (Lin et al. 2014). The high-attention tokens from the

[*]These authors contributed equally.

| Token Budget | Method | Venue | GQA | MMB | MMB$^{CN}$ | MME | POPE | SQA$^{IMG}$ | VQA$^{V2}$ | VQA$^{Text}$ | VizWiz | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 576 (100.0%) | LLaVA-1.5-7B | *CVPR'24* | 61.9 | 64.7 | 58.1 | 1862 | 85.9 | 69.5 | 78.5 | 58.2 | 50.0 | 100.0% |
| 192 (33.3%) | FastV | *ECCV'24* | 52.7 | 61.2 | 57.0 | 1613 | 64.8 | 67.3 | 67.1 | 52.5 | 50.8 | 90.4% |
| | HiRED† | *AAAI'25* | 58.8 | 62.6 | 54.5 | 1742 | 83.0 | 67.9 | 75.0 | - | 51.1 | 96.4% |
| | SparseVLM† | *ICML'25* | 59.5 | 64.1 | 58.0 | 1780 | 85.4 | 68.8 | 77.0 | 57.7 | 50.6 | 98.6% |
| | VisionZip | *CVPR'25* | 59.3 | 63.0 | - | 1783 | 85.3 | 68.9 | 76.8 | 57.3 | - | 97.7% |
| | PyramidDrop | *CVPR'25* | 57.3 | 63.3 | 56.8 | 1797 | 82.3 | 69.0 | 75.1 | 56.5 | 51.1 | 97.2% |
| | **HiPrune** | ***Ours*** | 59.2 | 62.8 | 57.0 | 1814 | 86.1 | 68.9 | 76.7 | 57.6 | 54.5 | **99.3%** |
| 128 (22.2%) | FastV | *ECCV'24* | 49.6 | 56.1 | 56.4 | 1490 | 59.6 | 60.2 | 61.8 | 50.6 | 51.3 | 85.4% |
| | HiRED† | *AAAI'25* | 57.1 | 61.7 | 53.9 | 1714 | 79.8 | 68.1 | 73.5 | - | 51.4 | 95.0% |
| | SparseVLM† | *ICML'25* | 53.8 | 64.4 | 58.1 | 1761 | 85.0 | 68.5 | 76.3 | 56.7 | 50.2 | 97.0% |
| | VisionZip | *CVPR'25* | 57.6 | 62.0 | - | 1762 | 83.2 | 68.9 | 75.6 | 56.8 | - | 96.2% |
| | PyramidDrop | *CVPR'25* | 57.1 | 61.6 | 56.6 | 1761 | 82.3 | 68.4 | 72.9 | 56.6 | 51.0 | 96.2% |
| | **HiPrune** | ***Ours*** | 57.3 | 62.2 | 56.4 | 1782 | 82.8 | 68.3 | 74.9 | 56.64 | 54.3 | **97.5%** |
| 64 (11.1%) | FastV | *ECCV'24* | 46.1 | 48.0 | 52.7 | 1356 | 48.0 | 51.1 | 55.0 | 47.8 | 50.8 | 76.7% |
| | HiRED† | *AAAI'25* | 54.6 | 60.2 | 51.3 | 1595 | 73.7 | 68.2 | 69.8 | - | 53.3 | 91.8% |
| | SparseVLM† | *ICML'25* | 53.7 | 60.1 | 52.5 | 1559 | 77.5 | 69.7 | 70.2 | 53.4 | 50.4 | 91.8% |
| | VisionZip | *CVPR'25* | 55.1 | 60.1 | - | 1690 | 77.0 | 69.0 | 72.4 | 55.5 | - | 92.7% |
| | PyramidDrop | *CVPR'25* | 47.5 | 58.8 | 50.5 | 1561 | 55.9 | 69.2 | 69.2 | 50.6 | 50.7 | 86.7% |
| | **HiPrune** | ***Ours*** | 53.6 | 59.5 | 53.4 | 1646 | 73.0 | 68.9 | 69.2 | 54.9 | 54.4 | **92.7%** |

Table 1: **Results on LLaVA-1.5-7B.** All methods are training-free. '†' denotes results reproduced by us.

| Layer | CLIP-L | CLIP-B | SigLIP | SigLIP2 | DeiT | VJEPA2 |
|---|---|---|---|---|---|---|
| 1 | 0.58× | 0.34× | 0.57× | 0.62× | 0.27× | 0.82× |
| *L/2* | **1**× | **1**× | **1**× | **1**× | **1**× | **1**× |
| *L* | 0.80× | 0.79× | 0.66× | 0.64× | 0.59× | 0.26× |

Table 2: **IoU of object segmentation mask and top 10% high-attention tokens.** Higher values stand for more overlap on objects in the image. '*L*' denotes the total layers in the encoder. The data is normalized for a better comparison.

middle layer share more overlap with objects than the input or output layer, indicating that **the attention from the middle layers is more correlated to objects in the image**.

Previous works have argued that high-attention tokens in the deep layer of ViT encode rich global information by conducting image classification tasks on these tokens (Darcet et al. 2024). In Fig. 1, the high-attention tokens in the output layer diffuse across the whole image and can serve as an ideal indicator of the image under a limited token budget. Therefore, we can conclude that **tokens receiving high attention in deep layers encode global information**.

### Prune Visual Tokens by Hierarchical Attention

Specifically, we first extract attention maps from a designated object layer $l$, selecting tokens with the highest attention scores as **Anchor Tokens**, which primarily correspond to object-centric regions. To enhance spatial continuity and suppress potential noise in attention estimation, we augment these anchors with adjacent **Buffer Tokens**. Together, anchor and buffer tokens encode detailed local semantics. The remaining token budget is allocated to **Register Tokens**, selected from the output layer based on attention scores. As shown in prior work (Darcet et al. 2024) and further validated in our study, these register tokens capture global contextual features, enabling a holistic representation of the image under tight token constraints.

## Experiment Results

We deploy HiPrune on LLaVA-1.5-7B and present comparison results in Table 1 with 192, 128, and 64 tokens retained. Across all the settings, HiPrune consistently outperforms existing methods, demonstrating superior performance. Specifically, with **1/3** tokens retained, HiPrune preserves **99.3%** of the original model's average performance, almost matching the vanilla model. Even under more constrained budgets, HiPrune maintains robust results, achieving **97.5%** with **128** tokens and **92.7%** with just **64** tokens.

## Conclusion and Future Direction

In this paper, we investigate the layer-wise attention patterns of vision encoders and reveal that middle layers predominantly capture object-centric features, while deeper layers emphasize global representations. Motivated by this insight, we propose HiPrune, a model-agnostic and training-free token pruning method for VLMs that leverages the hierarchical attention structure within the vision encoder. Future work will focus on applying HiPrune to more VLMs and more precise head-level analyses.

## Acknowledgements

This work is supported in part by the National Natural Science Foundation of China under grant 62571298. We are

grateful to Prof. Bin Chen and Yaowei Wang for their fruitful corrections and inspiration.

## References

Darcet, T.; Oquab, M.; Mairal, J.; and Bojanowski, P. 2024. Vision transformers need registers. In *ICLR*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *CVPR*, 26296–26306.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PMLR.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.