HAWK: Head Importance-Aware Visual Token Pruning in Multimodal Models

Supplementary Material

A. Extended Analysis on Visual Head Ablation

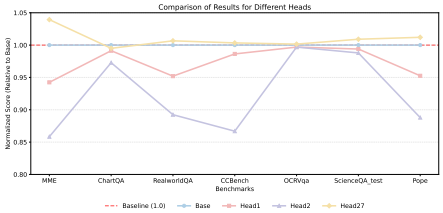
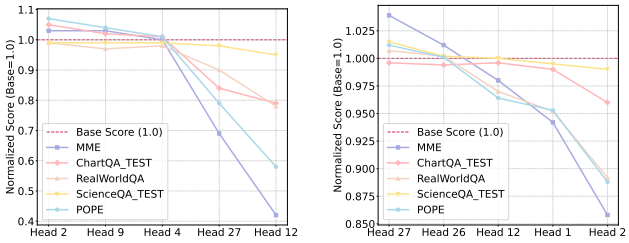


Figure 1. **Visual Head Ablation Study.** Comparison of results across different benchmarks relative to the Base model. The red dashed line indicates the baseline performance (1.0).

As discussed in the main text, masking the visibility of visual tokens for specific attention heads results in a consistent pattern of performance variation across diverse tasks. To further verify the generalizability of this finding, we expanded our evaluation scope in this section. Specifically, we selected three representative attention heads (Head 1, Head 2, and Head 27) for in-depth assessment across a broader suite of benchmarks, including MME[1], ChartQA[9], RealworldQA[13], CCBench[7], OCRVqa[10], ScienceQA-IMG[8], and POPE[6]. Note that certain benchmarks were excluded from this analysis due to Out-Of-Memory (OOM) errors encountered during evaluation to ensure experimental feasibility.

The results, illustrated in Figure 1, strongly corroborate our conclusions in the main text regarding the functional specificity of visual heads. Specifically, Head 2 demonstrates a critical role across the majority of benchmarks; masking this head causes substantial performance degradation, particularly on tasks demanding fine-grained visual perception such as CCBench and POPE (where normalized scores drop below 0.9). In contrast, masking Head 27 yields no negative impact and even leads to marginal performance gains on specific tasks, suggesting that this head likely encodes redundant information or visual noise. Furthermore, the varying sensitivity to head masking across benchmarks, exemplified by the drastic fluctuations on CCBench versus the relative stability on ScienceQA, further underscores the high dependency of complex reasoning tasks on specific critical visual heads. We validated head importance on **5 datasets** across **two models** (LLaVA, Qwen). As shown in **Fig. 2**, head rankings remain consistent across datasets for each model. This proves our method is **dataset-independent** and relies on intrinsic model features. We will **open-source** these universal head weights for mainstream MLLMs in the future.



(a) LLava 1.5-7b (b) Qwen 2.5vl-7b

Figure 2. Cross-model analysis of visual head ablation.

Table 1. Comparison with baselines at varying pruning ratios.

Method	Ratio	MME	ChartQA	TextVQA	AI2D	RealWorld	Avg. Rel.
Full Model	1.0	2315	86.2	85.2	80.7	67.7	100%
VisionZip	0.6	2308	78.0	75.3	79.9	69.8	96.1%
DART	0.6	2260	64.7	73.8	73.2	66.1	89.5%
HAWK	0.6	**2313**	**83.6**	**85.0**	**79.9**	67.6	**99.1%**
VisionZip	0.8	2182	62.1	70.1	77.9	68.2	89.2%
DART	0.8	2135	47.3	67.2	68.3	62.0	80.4%
HAWK	0.8	**2311**	**76.8**	**83.0**	**78.1**	65.0	**95.8%**
VisionZip	0.9	1923	43.5	61.4	71.0	63.7	77.5%
DART	0.9	1992	34.6	59.8	65.2	56.9	72.2%
HAWK	0.9	**2101**	**65.2**	**79.8**	**75.3**	60.4	**88.5%**

B. More results compared with other methods

We extended our evaluation to include comparisons with DART [12] and VisionZip [14]. As detailed in Table 1, HAWK consistently outperforms both baselines. Notably, at a 0.8 pruning ratio, HAWK preserves **95.8%** of the original performance, surpassing VisionZip (**89.2%**) and DART (**80.4%**) by substantial margins of **6.6%** and **15.4%**, respectively.

C. Computation Cost of Head Weights and Description of Benchmarks

The offline ablation is a negligible **one-time** cost. As shown in Table 2, calculating head importance takes only **0.39–0.83 hours** per dataset on 8 H20 GPUs. In contrast, training-based methods require hundreds of hours. Table3 provides a brief overview of the benchmarks.

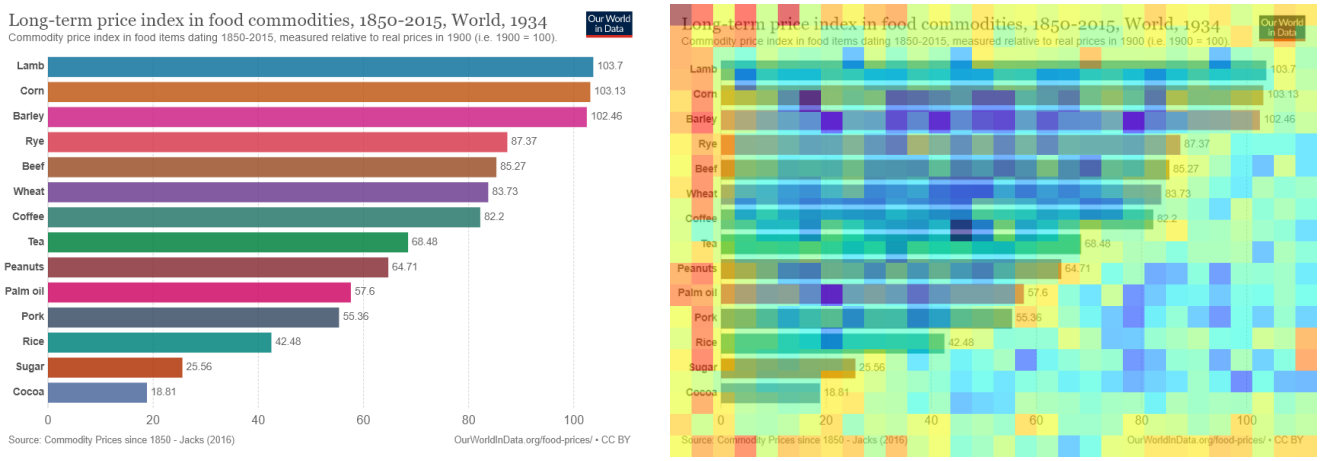
Table 2. Overhead of head weight calculation on LLaVA-1.5-7B

Dataset	**Time (hours)**	**GPU Memory (GB)**
MME	0.56	14.30
POPE	0.83	14.21
RealWorldQA	0.39	14.33

D. Additional Visualization Results

Benchmark	Task Type	Description
Image-based Benchmarks		
POPE [6]	Object Hallucination	Evaluates object hallucination ratios using random, popular, and adversarial sampling settings.
HallBench [3]	Visual Hallucination	Focuses on visual illusion and hallucination detection, requiring detailed image context reasoning.
MME [1]	Comprehensive Evaluation	A comprehensive suite covering 14 subtasks, including perception and cognition capabilities.
ScienceQA-IMG [8]	Science VQA	Contains multimodal science questions with annotated reasoning explanations and image contexts.
RealWorldQA [13]	Spatial Reasoning	Evaluates spatial perception and reasoning capabilities in diverse real-world environments.
CCBench [7]	Cultural Understanding	A benchmark designed to assess the model's understanding of Chinese cultural contexts and general knowledge.
TextVQA [11]	OCR-VQA	Requires reading and reasoning about text embedded in natural images to answer questions.
OCRVQA [10]	OCR-VQA	Focuses on visual question answering based on text-rich images such as book covers.
ChartQA [9]	Chart Understanding	Involves reasoning over complex charts and graphical data, requiring numerical and logical analysis.
AI2D [5]	Diagram Understanding	Evaluates the comprehension of science diagrams and textbook illustrations.
Video-based Benchmarks		
Video-MME [2]	Video Understanding	A comprehensive benchmark for long-duration video understanding across diverse domains.
WorldSense [4]	Video Reasoning	Assesses the understanding of physical laws, world dynamics, and causal reasoning in videos.

Table 3. **Overview of Evaluation Benchmarks.** We select ten image-based benchmarks (divided into general/reasoning and text/diagram categories) and two video-based benchmarks to comprehensively evaluate the model's capabilities.

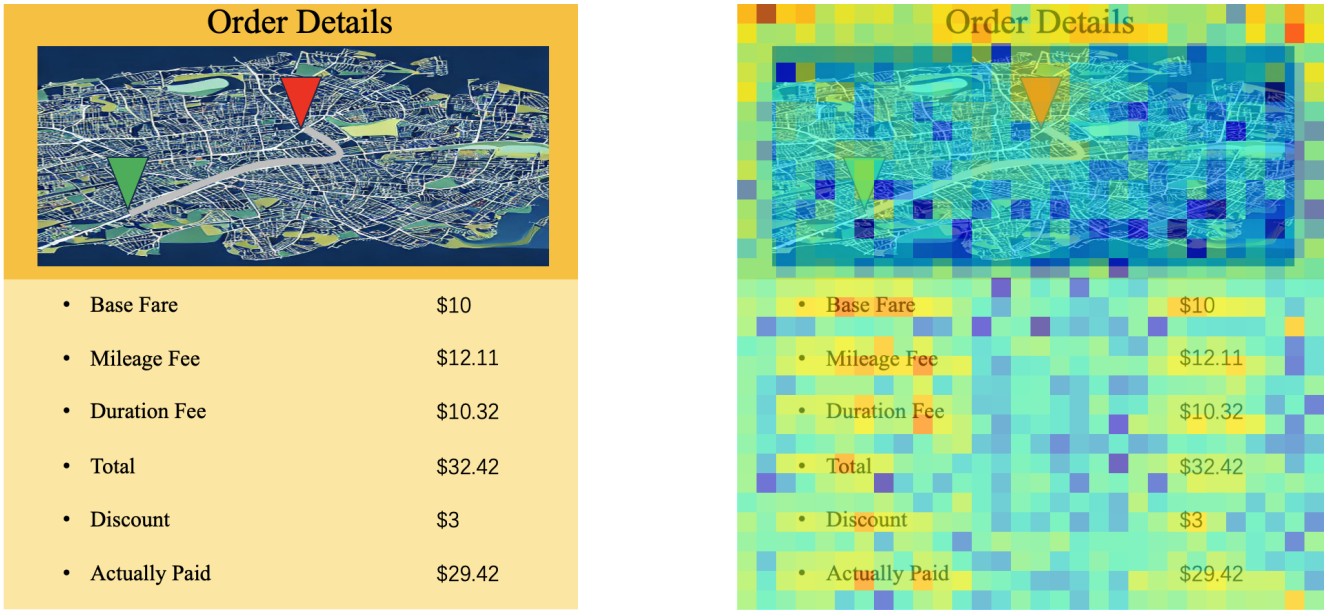


Example 1: ChartQA (Chart Understanding)

Query: *How many food item is shown in the bar graph?*

Divprune: The bar graph shows seven different food items: Lamb, Barley, Beef, Tea, Pork oil, Sugar, and Cocoa.
CDPruner: The bar graph shows six different food items. The items listed are: 1. Palm oil, 2. Pork, 3. Rice, 4. Sugar, 5. Cocoa, 6. Wheat.
HAWK (Ours): The bar graph shows **14** different food items. These are: Lamb, Corn, Barley, Rye, Beef, Wheat, Coffee, Tea,Peanuts, Palm oil, Pork, Rice, Sugar, Cocoa. So, there are 14 food items shown in the bar graph.

Figure 3. **Qualitative Results: Heatmaps and Response Comparison.** We present attention heatmaps to visualize the visual tokens retained by our HAWK, **where redder regions indicate higher attention scores**, alongside a qualitative comparison of the generated responses against other baseline methods. (Figure continued on the next page...)



Example 2: [MME/ Comprehensive Evaluation]

Query: *Here are the order details for my taxi ride. Should I actually pay $29.42? Please answer yes or no.*

Divprune: No.
CDPruner: No.
HAWK (Ours): **Yes.**

Example 3: [POPE/ Object Hallucination]

Query: *Is there a bottle in the image?*

Divprune: Yes,there is a bottle in the image. It appears to be a glass bottle with a label, located on the right side of the image, near the top.
CDPruner: No,there is no bottle in the image. The image shows a display of oranges and a sign advertising "Orange Juice" for sale. The sign mentions "A Small Bottle," but there are no actual bottles visible in the image.
HAWK (Ours): **Yes**,there is a bottle in the image. It is located on the right side of the image, near the top. The bottle appears to be part of the display for the orange juice stand.

Figure 3. Continued.

References

[1] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. 1, 2

[2] Chaoyou Fu, Yuhan Dai, Yongdong 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 2

[3] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024. 2

[4] Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. 2025. 2

[5] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer, 2016. 2

[6] 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, 2

[7] 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, 2

[8] 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, 2

[9] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279, 2022. 1, 2

[10] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019. 1, 2

[11] 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*, pages 8317–8326, 2019. 2

[12] Zichen Wen, Yifeng Gao, et al. Stop looking for important tokens in multimodal language models: Duplication matters more. *arXiv preprint arXiv:2502.11494*, 2025. 1

[13] xAI. Grok-1.5 vision preview. https://x.ai/blog/grok-1.5v, 2024. Introduces the RealWorldQA benchmark. Accessed: 2026-03-13. 1, 2

[14] Senqiao Yang, Yukang Chen, et al. Visionzip: Longer is better but not necessary in vision language models. *arXiv preprint arXiv:2412.04467*, 2024. 1