Reasoning with Fewer Eyes: Efficient Visual Token Withdrawal for Multimodal Reasoning

Andrea Ramazzina[1,2] **Tobias Haab**[1] **David Fitzek**[3] **Stefano Gasperini**[2,4]
Jonas Uhrig[1] **Mario Bijelic**[5]
[1]Mercedes-Benz AG [2] Technical University of Munich [3] Fraunhofer IAIS
[4] Munich Center for Machine Learning [5] Princeton University
{andrea.ramazzina, tobias.haab, jonas.uhrig}@mercedes-benz.com
david.fitzek@iais.fraunhofer.de stefano.gasperini@tum.de
mario.bijelic@princeton.edu

Abstract

Vision-Language models have shown strong promise for multimodal reasoning tasks, where autoregressive generation allows the model to combine perception and abstract reasoning. However, especially when processing high-resolution images or long videos, the large number of visual tokens severely slows down inference. Drawing from the observation that attention devoted to vision tokens consistently drops during autoregressive text generation, we propose a simple method to accelerate multimodal reasoning: after the model has generated a small number of text tokens, we remove all vision tokens from subsequent decoding steps. This reduces both memory usage and computation, while retaining the model's ability to ground its reasoning in the visual input. Our approach requires no additional training and is fully compatible with popular efficiency techniques such as KV caching and FlashAttention. Experiments on multiple datasets and with different models demonstrate that our method achieves substantial speedups with minimal impact on reasoning accuracy.

1 Introduction

Large language models have driven substantial advances in language understanding, reasoning, and text generation [1, 3, 29]. Building on this foundation, multimodal reasoning models integrate visual encoders such as CLIP [22] to enable multimodal understanding. These models demonstrate strong performance on multimodal reasoning tasks through autoregressive architectures that effectively combine visual perception with abstract reasoning [12, 14, 2, 13, 4].

However, these reasoning tasks are computationally expensive, with visual tokens increasing the burden, especially for high-resolution images or extended sequences. High-resolution images generate hundreds to thousands of visual tokens, from 576 tokens in LLaVA [17] to over 2000 for larger inputs [32]. This creates significant computational overhead, and thus severely impacting inference speed for complex reasoning tasks.

To address this, recent methods attempt to reduce computational costs through visual token reduction [25, 7], but these approaches were primarily designed for perception-focused tasks rather than reasoning scenarios. These approaches identify "important" tokens based on predefined metrics and prune or merge the remainder. However, they suffer from several limitations that hinder practical application and deployment for real-world open-domain use cases.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Efficient Reasoning.

Rather than designing complex heuristics or learning-based methods, we first examine whether visual tokens remain equally important throughout the reasoning process. We find that attention to visual tokens consistently drops during autoregressive text generation, suggesting their importance diminishes as models transition from perception to abstract reasoning. Based on this observation, we propose a training-free method that removes visual tokens after generating a small number of text tokens. Our approach requires no retraining, introduces no parameters, and is fully compatible with KV caching [20] and FlashAttention [8].

We evaluate our method on two multimodal mathematical reasoning benchmarks: MathVerse[24], which focuses on visually grounded mathematical reasoning, and WeMath[21], designed for step-by-step multimodal math problem solving. Our approach achieves up to 56% speedup with minimal accuracy degradation, proving most effective for complex reasoning sequences where efficiency is critical. The method offers three key advantages: substantial reduction in inference time and memory usage, training-free compatibility with existing models, and full compatibility with other efficiency techniques for compound acceleration benefits.

2 Related Work

2.1 Multimodal reasoning models

Multimodal reasoning models have evolved toward advanced multi-step inference capabilities. Foundational models like LLaVA [17] and BLIP-2 [13] established visual-textual integration paradigms through feature projection and trainable querying mechanisms, respectively, enabling competent performance on standard reasoning tasks but often struggling with complex multi-step scenarios.

Recent advances emphasize structured reasoning and RL enhancement. LLaVA-CoT [31] decomposes problems into explicit stages (Summary, Caption, Reasoning, Conclusion) for systematic inference, while RL-enhanced models like MM-EUREKA [19], VLM-R1 [27], DeepSeek-R1 [11], and QwQ [34] leverage GRPO [26] and training techniques for mathematical and long-horizon reasoning. Kimi-VL [9] combines architectural innovations with mixture-of-expert frameworks for efficient complex reasoning.

2.2 Vision token reduction

Token reduction methods decrease computational costs through pruning [23, 33, 15] and merging [5, 18] strategies. Early approaches like EViT [15] and Evo-ViT [33] fuse non-critical tokens, while ToMe [5] employs soft-matching algorithms. In multimodal settings, CrossGET [28] and MADTP [6] use cross-modal alignment tokens to guide reduction, LLaVA-PruMerge [25] leverages spatial redundancy, and FastV [7] removes half the tokens after early decoder layers.

More recently, Visual Token Withdrawal (VTW) [16] withdraws vision tokens at specific decoder layers based on KL divergence criteria. However, these layer-based approaches maintain vision tokens throughout substantial portions of the reasoning process, particularly during the critical transition from visual grounding to abstract reasoning where efficiency is most needed.

3 Method

This section describes our method, referred as MVW (M-step Vision Withdrawal) for efficient multimodal reasoning. We first motivate the approach with empirical evidence showing the influence of attention to vision tokens during autoregressive decoding and then present our method.

3.1 Intuition and empirical motivation

A key motivation in our proposed MVW approach lies in the visual attention pattern that we observe happening in different multimodal reasoning models.

Specifically, during autoregressive generation, multimodal reasoning decoders attend strongly to vision tokens in the initial steps, using them to ground the reasoning in the visual input. As the sequence grows, however, attention to vision tokens steadily declines, and the model increasingly

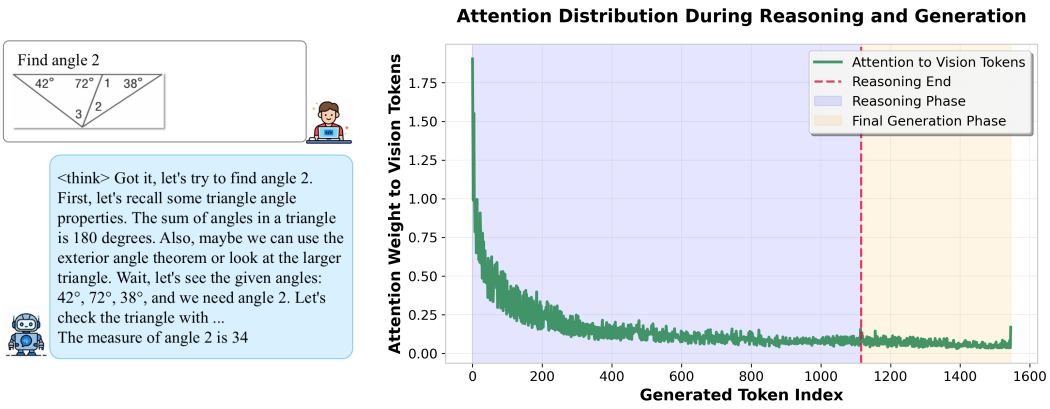

Figure 1: Left: possible input to a reasoning VLM with image and text prompt with subsequent reasoning. Right: the total attention assigned to vision tokens as a function of generated token index is shown. The attention to vision tokens drops rapidly after the first few tokens.

relies on abstract reasoning over the previously generated text. This behavior is in line with the attention sink phenomenon described in [30] and followup works [16].

We illustrate this phenomenon in Fig. 1, which plots the total attention mass assigned to vision tokens at each generation step for the reasoning model Kimi-VL [10]. The curve shows high attention early on, followed by a consistent decrease that eventually stabilizes at around 5% of the initial value, indicating that prolonged access to vision tokens is largely unnecessary once the reasoning phase progresses. This behavior persists also after the reasoning ends, indicating that the model does not rely on the input image to generate the final answer.

3.2 M-step Vision Withdrawal

In this section, we first outline the standard inference process of reasoning VLMs, and then introduce how vision token withdrawal can be incorporated into these models.

Given a system prompt, an image or video, and an instruction prompt, a VLM model first individually processes them to obtain respectively a set of system tokens S, vision tokens V and instruction tokens I. Subsequently, a decoder transformer D is usually employed to generate the next token O_t autoregressively, that is:

$$O_{t+1} = D(S, V, I, O_{1..t}) \tag{1}$$

This process is repeated until an end-of-sequence (`<eos>`) token is produced, or a predefined maximum generation length is reached. The final answer is then obtained by concatenating and detokenizing the sequence of generated tokens $O_{1,..,t} = \{O_1, \ldots, O_t\}$ into natural language text.

Depending on the architecture, the vision tokens V may be integrated into the decoding process in different ways: some models employ dedicated cross-attention layers between text and vision streams, while others project V into the language embedding space and prepend them to the textual input, such that they are handled entirely through self-attention.

Reasoning-oriented VLMs follow the same general decoding process, but are typically trained to first produce intermediate reasoning tokens (often enclosed in special delimiters such as `<think>` ... `</think>`) before generating the final answer tokens. This allows the model to externalize multi-step reasoning while still adhering to the standard autoregressive generation framework.

In our proposed modification, the decoding process proceeds identically to the standard formulation for the first M generated tokens or until the reasoning process has ended. After either of these two conditions is met, decoding is performed without the vision token, that is:

$$O_{t+1} = \begin{cases} D(S, I, O_{1..t}) & \text{if } t > M \vee R \in O_{1..t} \\ D(S, V, I, O_{1..t}) & else \end{cases} \tag{2}$$

Table 1: Performance comparison across different methods on the MathVerse and WeMath datasets. Best results are in **bold**, second-best are underlined.

Model	Method	TFLOPs↓	Accuracy↑	
			MathVerse [24]	**WeMath** [21]
LLaVA-CoT [31]	Baseline	494.8 (100%)	44.2	**50.86**
(Llama-3.2-11B-Vision)	VTW[16] (K=20)	394.2 (79.7%)	37.34	40.69
	VTW[16] (K=30)	444.5 (89.8%)	42.02	50.46
	Ours (M=5)	**382.1 (77.2%)**	**44.59**	50.00
	Ours (M=10)	389.0 (78.6%)	43.76	50.63
Kimi-VL [10]	Baseline	434.2 (100%)	**70.27**	68.56
(Kimi-VL-A3B-Thinking-2506)	VTW[16] (K=13)	238.9 (55.0%)	52.47	53.56
	VTW[16] (K=20)	336.6 (77.5%)	70.09	67.47
	Ours (M=500)	**191.5 (44.1%)**	67.61	67.01
	Ours (M=750)	258.5 (59.5%)	67.66	**69.65**

where R represents the end-of-reasoning token (e.g. `</think>` for Kimi-VL [10]). Note that the counting of M begins after the reasoning segment is initiated (e.g., following the `<REASONING>` tag), which may not coincide with the beginning of the output sequence in models such as LLaVA-CoT. In practice, how the vision tokens are withdrawn depends on the model architecture :

1. **For Models using Self-Attention,** withdrawal is realized by simply removing all prepended vision tokens from the input sequence while leaving the rest of the decoding pipeline unchanged. Importantly, positional encodings and subsequent autoregressive processing remain unaffected.
2. **For Models using Cross-Attention,** such as LLaMA 3.2-based VLMs, vision withdrawal is implemented by skipping the cross-attention layers altogether once the withdrawal condition is met. That is, after M tokens or at the end-of-reasoning marker R, the decoder reverts to pure self-attention updates with no contribution from vision features.

4 Experiments & Results

We evaluate our approach on two recently proposed multimodal reasoning benchmarks: **MathVerse** [24], which focuses on visually grounded mathematical reasoning, and **WeMath** [21], a dataset designed for assessing step-by-step multimodal math problem solving.

For models, we consider two representative reasoning VLMs. The first is **LLaVA-CoT** [31], a chain-of-thought finetuned variant of LLaMA-3.2-Vision that follows a vision cross-attention design. The second is **Kimi-VL** [10], a native multimodal reasoning LLM that integrates image features directly into the self-attention stream, enabling joint reasoning over visual and textual tokens.

We benchmark three settings for each model: the baseline without any efficiency intervention, Vision Token Withdrawal (VTW) [16] where vision tokens are dropped at different decoder layers ($K = 20$ and $K = 30$), and our method, which removes all vision tokens after the generation of M tokens, with different cutoffs, namely $M = 500/750$ for Kimi-VL and $M = 5/10$ for LLaVA-CoT (lower as the reasoning is preceded by the summary and captioning phases).

We report the results in Table 1. On both MathVerse and WeMath, our method consistently matches or outperforms the baselines in terms of reasoning accuracy, while requiring substantially fewer FLOPs.

These results demonstrate that removing visual tokens at mid-late stage of the reasoning phase preserves reasoning quality while accelerating decoding, confirming the intuition that prolonged visual access is unnecessary for deep reasoning.

5 Conclusion

We propose MVW, a vision token withdrawal approach to accelerate multimodal reasoning with minimal accuracy loss, while being fully compatible with KV caching and FlashAttention. Our

method requires no retraining and applies across different VLM architectures, offering a practical drop-in solution for efficient large-scale vision-language inference.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, et al. GPT-4 Technical Report. *ArXiv e-prints*, March 2023.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, et al. Flamingo: a Visual Language Model for Few-Shot Learning. *ArXiv e-prints*, April 2022.

[3] Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, et al. Gemini: A Family of Highly Capable Multimodal Models. *ArXiv e-prints*, December 2023.

[4] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, et al. PaliGemma: A versatile 3B VLM for transfer. *ArXiv e-prints*, July 2024.

[5] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *ICLR*, pages 1–20, 2023.

[6] Jianjian Cao, Peng Ye, Shengze Li, Chong Yu, Yansong Tang, Jiwen Lu, and Tao Chen. Madtp: Multimodal alignment-guided dynamic token pruning for accelerating vision-language transformer, 2024.

[7] Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, et al. An Image is Worth 1/2 Tokens After Layer 2: Plug-and-Play Inference Acceleration for Large Vision-Language Models. *ArXiv e-prints*, March 2024.

[8] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

[9] Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, et al. Kimi-VL Technical Report. *ArXiv e-prints*, April 2025.

[10] Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, et al. Kimi-VL technical report, 2025.

[11] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *ArXiv e-prints*, January 2025.

[12] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, et al. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. *ArXiv e-prints*, July 2024.

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR, July 2023.

[14] Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, et al. Perception, Reason, Think, and Plan: A Survey on Large Multimodal Reasoning Models. *ArXiv e-prints*, May 2025.

[15] Youwei Liang, GE Chongjian, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Evit: Expediting vision transformers via token reorganizations. In *ICLR*, pages 1–21, 2022.

[16] Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5334–5342, 2025.

[17] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *ArXiv e-prints*, April 2023.

[18] Dmitrii Marin, Jen-Hao Rick Chang, Anurag Ranjan, Anish Prabhu, Mohammad Rastegari, and Oncel Tuzel. Token pooling in vision transformers, 2021.

[19] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, et al. MM-Eureka: Exploring the Frontiers of Multimodal Reasoning with Rule-based Reinforcement Learning. *ArXiv e-prints*, March 2025.

[20] Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Jonathan Heek, Kefan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference. In *MLSys*, 2023.

[21] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, et al. Learning Transferable Visual Models From Natural Language Supervision. *ArXiv e-prints*, February 2021.

[23] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *NeurIPS*, pages 13937–13949, 2021.

[24] Yichi Zhang Haokun Lin Ziyu Guo Pengshuo Qiu Aojun Zhou Pan Lu Kai-Wei Chang Peng Gao Hongsheng Li Renrui Zhang, Dongzhi Jiang. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *arXiv*, 2024.

[25] Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive Token Reduction for Efficient Large Multimodal Models. *ArXiv e-prints*, March 2024.

[26] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, et al. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *ArXiv e-prints*, February 2024.

[27] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, et al. VLM-R1: A Stable and Generalizable R1-style Large Vision-Language Model. *ArXiv e-prints*, April 2025.

[28] Dachuan Shi, Chaofan Tao, Anyi Rao, Zhendong Yang, Chun Yuan, and Jiaqi Wang. Crossget: Cross-guided ensemble of tokens for accelerating vision-language transformers, 2023.

[29] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *ArXiv e-prints*, July 2023.

[30] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

[31] Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[32] Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, et al. LLaVA-UHD: an LMM Perceiving Any Aspect Ratio and High-Resolution Images. *ArXiv e-prints*, March 2024.

[33] Yifan Xu, Zhijie Zhang, Mengdan Zhang, Kekai Sheng, Ke Li, Weiming Dong, Liqing Zhang, Changsheng Xu, and Xing Sun. Evo-vit: Slow-fast token evolution for dynamic vision transformer. In *AAAI*, pages 2964–2972, 2022.

[34] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 Technical Report. *ArXiv e-prints*, May 2025.