Learnable vs. Heuristic Visual Token Compression in Cross-Domain Visual-Language Benchmarks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does the performance of learnable visual token compression techniques compare to heuristic-based methods in cross-domain visual-language benchmarks like VQAv2 and COCO-QA, measured by accuracy. 9 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.9/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: VisionSelector: End-to-End Learnable Visual Token Compression for Efficient Multimodal LLMs. Research question: How does the performance of learnable visual token compression techniques compare to heuristic-based methods in cross-domain visual-language benchmarks like VQAv2 and COCO-QA, measured by accuracy and FLOPs efficiency?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.9/10.

3 Results

11 papers retrieved. 9 claims extracted; 0 independently verified. Quality review score: 3.9/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
VisionSelector outperforms prior SOTA under a 10% token budget on Qwen2.5-VL-7B, retaining 87.75% performance on average	×	0.05
VisionSelector achieves a 2$\times$ prefill speedup with only 5% tokens on DocVQA.	×	0.04
VisionSelector delivers a leading three-way trade-off in accuracy, GPU memory, and speedup on DocVQA with a 20% token bu	×	0.05
VisionSelector improves overall performance by 12.14 percentage points at a 10% token retention.	×	0.09
VisionSelector accelerates the prefill phase by a factor of 1.73$\times$ while reducing memory consumption to 86.08% at 20% ret	×	0.03
VisionSelector requires only 12.85M trainable parameters and approximately 40 minutes of training on 8 A800 NVIDIA GPUs.	×	0.09
VisionSelector achieves 89.91% Anls, 770 Relaxed, 49.22% Acc, 2293.54 Score, and 84.27 F1 on DocVQA, ChartQA, MMMU, MME,	×	0.02
VisionSelector achieves 66.55% Acc, 59.82% Acc, 59.22% Score, and 27.10% WUPS on MVBench, SEEDBench, VideoMME, and NextQ	×	0.02
VisionSelector reduces Max GPU memory to 17.57GB and E2E Latency to 924.57ms on MVBench, SEEDBench, VideoMME, and NextQA	×	0.01

References

- http://arxiv.org/abs/2510.16598v1

- http://arxiv.org/abs/2412.16117v1
- http://arxiv.org/abs/2510.07143v3