Entropy Hypothesis Generalization in Multimodal Models Across Cross-Domain Benchmarks

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 19 peer-reviewed papers addressing the following research question: Does the ENTROPY hypothesis (initial image size reduction) generalize to multimodal models (e.g., visual-language models like CLIP) when evaluating performance on cross-domain benchmarks (e.g., VCR. 18 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.8/10. 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: Does the ENTROPY hypothesis (initial image size reduction) generalize to multimodal models (e.g., visual-language models like CLIP) when evaluating performance on cross-domain benchmarks (e.g., VCR vs. Flickr30k), as measured by accuracy and FLOPs?.

2 Methodology

Systematic literature search across multiple databases yielded 19 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.8/10.

3 Results

19 papers retrieved. 18 claims extracted; 0 independently verified. Quality review score: 3.8/10.

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 retains 87.75% of the baseline performance on average under a 10% token budget on Qwen2.5-VL-7B.	×	0.05
VisionSelector achieves a 2x prefill speedup on DocVQA with only 5% token retention.	×	0.05
VisionSelector requires 12.85M trainable parameters.	×	0.13
Training VisionSelector takes approximately 40 minutes on 8 A800 NVIDIA GPUs.	×	0.03
At 10% token retention, VisionSelector improves overall performance by 12.14 percentage points compared to the baseline.	×	0.08
At 20% token retention, VisionSelector accelerates the prefill phase by a factor of 1.73x.	×	0.04
At 20% token retention, VisionSelector reduces memory consumption to 86.08% of the baseline.	×	0.04
The baseline Qwen2.5-VL-7B model uses an average of 1951.61 visual tokens for DocVQA.	×	0.04
The baseline Qwen2.5-VL-7B model achieves a DocVQA Anls score of 94.33.	×	0.03
VisionSelector achieves a DocVQA Anls score of 89.91 with a 10% token retention configuration.	×	0.04
VisionSelector achieves an MME Score of 2293.54 in the reported configuration.	×	0.02
VisionSelector achieves a POPE F1 score of 84.27 in the reported configuration.	×	0.01
VisionSelector achieves an average accuracy of 95.96% relative to the upper bound in the configuration listed in Table (	×	0.02
VisionSelector achieves a Max GPU memory usage of 17.57 GB on the MVBench benchmark.	×	0.02
VisionSelector achieves a Prefill Time of 760.82 ms on the MVBench benchmark.	×	0.02
VisionSelector achieves an End-to-End Latency of 924.57 ms on the MVBench benchmark.	×	0.07
VisionSelector comprises a Differentiable Top-K Selection Mechanism, a Curriculum Annealing Strategy, and a backbone-dec	×	0.14
The Learnable Importance Scorer (LIS) computes global token importance within a single forward pass.	×	0.04

References

- https://arxiv.org/abs/2502.03950
- https://arxiv.org/abs/2510.16598
- https://arxiv.org/abs/2503.20322