Learned Attention Pruning Patterns Boost ViT Inference Throughput in Code Generation

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: What is the impact of self-attention mechanism complexity reduction techniques on the inference throughput of ViTs in code generation tasks, measured by tokens per second on MBPP. 12 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Which Tokens to Use? Investigating Token Reduction in Vision Transformers. Research question: What is the impact of self-attention mechanism complexity reduction techniques on the inference throughput of ViTs in code generation tasks, measured by tokens per second on MBPP?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.7/10.

3 Results

15 papers retrieved. 12 claims extracted; 2 independently verified. Quality review score: 4.7/10.

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
Pruning-based methods with learned reduction patterns are consistently among the top-3 methods across all datasets.	✓	0.27
Soft-clustering methods are consistently among the bottom three methods across all datasets.	×	0.07
With the DeiT-T and DeiT-S backbones, hard-merging methods ToMe and DPC-KNN regularly outperform all other methods, espe	×	0.03
With the DeiT-B backbone, pruning-based methods with learned reduction patterns outperform even the hard-merging methods	✓	0.20
Fixed-pattern p methods are competitive when 90% of tokens are kept, but at lower keep rates the performance drops sign	×	0.05
DynamicViT method is the most unstable of the tested methods, often being in the bottom three methods when the keep rate	×	0.03
ATS method has great performance on the challenging NUS-WIDE dataset but average performance on all other datasets.	×	0.06
ATS method uses on average 50-90 and 10-30 fewer tokens than the other methods.	×	0.05
Orthogonal Procrustes distance and NMI are highly correlated with the difference in model performance.	×	0.04
IoU metric is moderately correlated with the difference in model performance.	×	0.04
NMI is a better proxy than IoU for model performance.	×	0.05
Top-K pruning method and its extension, EViT, are found to be the best performing methods.	×	0.06

References

- http://arxiv.org/abs/2209.15001v3
- http://arxiv.org/abs/2303.15105v1
- http://arxiv.org/abs/2308.04657v1