SOVEREIGN: What is the impact of dynamic token count on FLOPs efficiency and reasoning accuracy when processing variable-

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Vision Transformers (ViTs) have achieved state-of-the-art performance across various computer vision tasks, but their high computational cost remains a challenge. Token pruning has been proposed to reduce this cost by selectively removing less important tokens. While effective in vision tasks by discarding non-object regions, applying this technique to audio tasks presents unique challenges, as distinguishing relevant from irrelevant regions in time-frequency representations is less straightforward. In this study, for the first time, we applied token pruning to ViT-based audio classification m

1 Introduction

Analysis of: Token Pruning in Audio Transformers: Optimizing Performance and Decoding Patch Importance. Research goal: What is the impact of dynamic token count on FLOPs efficiency and reasoning accuracy when processing variable-complexity images with different tokenization strategies?.

2 Methodology

Multi-query arXiv search (4 parallel queries, Relevance-sorted). TF-IDF cosine semantic verification (bigrams, threshold=0.15). NIM nv-embedqa-e5-v5 (dim=1024) for semantic indexing. Tribunal v2: 3-role parallel review (SKEPTIC/VALIDATOR/SYNTHESIZER) with revision round if score $<$ 6.5.

3 Results

12 papers retrieved. 8 claims extracted, 8 verified. Tribunal: 7.5/10 $\rightarrow$ APPROVE (revision_round=0). Policy: AUTO_APPROVE.

4 Uncertainties

NIM free tier latency varies. TF-IDF verification is a weak signal. arXiv Relevance ranking is query-dependent. Tribunal consensus is LLM-based and prompt-sensitive.

5 Extracted Claims

Claim	Verified	Confidence
TopK token pruning can reduce MAC operations of AudioMAE and AST by 30-40%, with less than a 1% drop in accuracy.	✓	0.33
High-intensity or high-variation tokens contribute significantly to model accuracy.	✓	0.27
Low-intensity or low variation tokens also remain important when token pruning is applied.	✓	0.31
Pruning solely based on the intensity or variation of signals in a patch leads to a noticeable drop in accuracy.	✓	0.30
There is a high correlation between attention scores and statistical features (intensity/variation) of tokens.	✓	0.24
Retained tokens consistently receive distinct attention compared to pruned ones.	✓	0.23
AudioMAE retains more low-intensity tokens than AST.	✓	0.22
AudioMAE's self-supervised reconstruction objective explains why it retains more low-intensity tokens than AST.	✓	0.23

References

- `http://arxiv.org/abs/2303.15105v1`
- `http://arxiv.org/abs/2605.23892v1`
- `http://arxiv.org/abs/2504.01690v2`