SOVEREIGN: How does dynamic token pruning in audio Transformers affect FLOPs efficiency and word error rate on the LibriS

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 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: How does dynamic token pruning in audio Transformers affect FLOPs efficiency and word error rate on the LibriSpeech benchmark compared to uniform token retention 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

10 papers retrieved. 11 claims extracted, 2 verified. Tribunal: 2.8/10 $\rightarrow$ REJECT (revision_round=0). Policy: ESCALATE_TO_OWNER.

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.	✓	0.30
AudioMAE retains more low-intensity tokens than AST.	✓	0.25
AudioMAE is more sensitive to token loss than AST especially at lower keep-rates.	×	0.05
Accuracy tends to drop more when more tokens are pruned—particularly for harder tasks with a larger number of classes.	×	0.06
AudioMAE's performance degrades more than AST's as the keep-rate decreases.	×	0.05
For AS-20K, TopK pruning at keep-rate 0.9 achieves 86.0% MAC reduction with 38.7 accuracy (0.0 drop).	×	0.04
For ESC-50, TopK pruning at keep-rate 0.8 achieves 74.9% MAC reduction with 94.32 accuracy (0.73 drop).	×	0.04
For SPC-2, TopK pruning at keep-rate 0.7 achieves 66.1% MAC reduction with 97.19 accuracy (0.14 drop).	×	0.04
For SPC-2, TopK pruning at keep-rate 0.4 achieves 44.6% MAC reduction with 97.07 accuracy (0.26 drop).	×	0.04
For AS-20K, TopK pruning at keep-rate 0.8 achieves 74.1% MAC reduction with 37.9 accuracy (0.8 drop).	×	0.04
For ESC-50, TopK pruning at keep-rate 0.6 achieves 56.7% MAC reduction with 94.37 accuracy (0.68 drop).	×	0.04

References

- `http://arxiv.org/abs/2504.01690v2`

- http://arxiv.org/abs/1912.05946v2
- http://arxiv.org/abs/2304.00649v1