SOVEREIGN: Does applying cross-modal token pruning from vision to audio Transformers degrade robustness to noise perturba

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: Does applying cross-modal token pruning from vision to audio Transformers degrade robustness to noise perturbations (e.g., on ESC-50) when optimizing for throughput on variable-length spectrograms?.

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. 7 claims extracted, 3 verified. Tribunal: 6.5/10 → REVISE (revision_round=1). 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.	✓	0.32
AudioMAE is more sensitive to token loss than AST especially at lower keep-rates.	×	0.05
AudioMAE's performance degrades more than AST's as the keep-rate decreases.	×	0.04
FastAST with ToMe achieves 88.4% MAC reduction with -0.6 accuracy drop on AS-20K dataset.	×	0.04
TopK pruning achieves 86.0% MAC reduction with 0.0 accuracy drop on AS-20K dataset.	×	0.07
Low-intensity or low-variation tokens remain important when token pruning is applied.	✓	0.34
AudioMAE retains more low-intensity tokens than AST.	✓	0.24

References

- http://arxiv.org/abs/2504.01690v2
- http://arxiv.org/abs/2211.13189v2
- http://arxiv.org/abs/2110.09784v2