SOVEREIGN: How does the inference throughput and token count scaling of adaptive token pruning (CAT-like) methods compare

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

In this paper, we introduce PruneVid, a visual token pruning method designed to enhance the efficiency of multi-modal video understanding. Large Language Models (LLMs) have shown promising performance in video tasks due to their extended capabilities in comprehending visual modalities. However, the substantial redundancy in video data presents significant computational challenges for LLMs. To address this issue, we introduce a training-free method that 1) minimizes video redundancy by merging spatial-temporal tokens, and 2) leverages LLMs' reasoning capabilities to selectively prune visual fea

1 Introduction

Analysis of: PruneVid: Visual Token Pruning for Efficient Video Large Language Models. Research goal: How does the inference throughput and token count scaling of adaptive token pruning (CAT-like) methods compare to fixed tokenization in video diffusion transformers across varying scene complexity, measured in FPS and latency on 4K resolution benchmarks?.

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. 4 claims extracted, 0 verified. Tribunal: 3.5/10 → 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
PruneVid achieves 42.6% visual token reduction on average for the PLLaVA model when using token merging.	×	0.10
PruneVid achieves 45.6% visual token reduction on average for the ST-LLM model when using token merging.	×	0.09
	×	0.00
PruneVid reduces visual tokens by 16.2% on average for the LLaVA-OneVision model.	×	0.07

References

- http://arxiv.org/abs/2504.01690v2
- http://arxiv.org/abs/2412.16117v1
- http://arxiv.org/abs/2507.07995v1