SOVEREIGN: What is the impact of token-level guided routing on inference latency and cross-modal reasoning accuracy in Mo

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Abstract In the past years, multimodal large language models (MLLMs) have demonstrated remarkable performance in tasks such as visual question answering and visual understanding and reasoning. However, the extensive model size and high training and inference costs have hindered the widespread application of MLLMs in academia and industry. Thus, studying efficient and lightweight MLLMs has enormous potential, especially in edge computing scenarios. In this survey, we provide a comprehensive and systematic review of the current state of efficient MLLMs. Specifically, this survey summarizes the t

1 Introduction

Analysis of: Efficient multimodal large language models: a survey. Research goal: What is the impact of token-level guided routing on inference latency and cross-modal reasoning accuracy in MoE vision-language models compared to dense baselines on the MMBench and SEED-Bench benchmarks?.

2 Methodology

Multi-query arXiv search (1 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

3 papers retrieved. 3 claims extracted, 3 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
Multimodal large language models (MLLMs) have demonstrated remarkable performance in tasks such as visual question answe	✓	0.41
The extensive model size and high training and inference costs have hindered the widespread application of MLLMs in acad	✓	0.37
Studying efficient and lightweight MLLMs has enormous potential, especially in edge computing scenarios	✓	0.35

References

- https://doi.org/10.1007/s44267-025-00099-6
- https://openalex.org/W7148218130
- https://doi.org/10.36227/techrxiv.176620829.92520878/v2