SOVEREIGN: What is the inference throughput (tokens per second) of Gemini 1.5 Flash versus LLaVA-NeXT on the Video-MME be

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Large foundation models, including large language models (LLMs), vision transformers (ViTs), diffusion, and LLM-based multimodal models, are revolutionizing the entire machine learning lifecycle, from training to deployment. However, the substantial advancements in versatility and performance these models offer come at a significant cost in terms of hardware resources. To support the growth of these large models in a scalable and environmentally sustainable way, there has been a considerable focus on developing resource-efficient strategies. This survey delves into the critical importance of s

1 Introduction

Analysis of: A Survey of Resource-efficient LLM and Multimodal Foundation Models. Research goal: What is the inference throughput (tokens per second) of Gemini 1.5 Flash versus LLaVA-NeXT on the Video-MME benchmark under single-GPU memory constraints?.

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. 6 claims extracted, 6 verified. Tribunal: 8.0/10 $\rightarrow$ 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
Large foundation models include large language models (LLMs), vision transformers (ViTs), diffusion models, and LLM-base	✓	0.28
Large foundation models are revolutionizing the machine learning lifecycle from training to deployment.	✓	0.26
Advancements in versatility and performance of large foundation models come at a significant cost in terms of hardware r	✓	0.29
There is a considerable focus on developing resource-efficient strategies to support the scalable and environmentally su	✓	0.30
The survey examines both algorithmic and systemic aspects of resource-efficient strategies for large foundation models.	✓	0.25
The survey covers topics including model architectures, training/serving algorithms, and practical system designs and im	✓	0.24

References

- https://doi.org/10.48550/arxiv.2401.08092
- https://doi.org/10.1038/s41467-025-61040-5
- https://doi.org/10.1609/aaai.v39i5.32567