SOVEREIGN: Scaling Laws for Native Multimodal Models

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Building general-purpose models that can effectively perceive the world through multimodal signals has been a long-standing goal. Current approaches involve integrating separately pre-trained components, such as connecting vision encoders to LLMs and continuing multimodal training. While such approaches exhibit remarkable sample efficiency, it remains an open question whether such late-fusion architectures are inherently superior. In this work, we revisit the architectural design of native multimodal models (NMMs)-those trained from the ground up on all modali-ties-and conduct an extensive sca

1 Introduction

Analysis of: Scaling Laws for Native Multimodal Models. Research goal: Does SMoES's soft modality-guided routing improve MoE-VLM accuracy on the MMMU benchmark compared to dense models of equivalent total parameter count, and how does this gap change when scaling from 7B to 34B total parameters?.

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

11 papers retrieved. 5 claims extracted, 5 verified. Tribunal: 7.7/10 $\rightarrow$ 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
The study spans 457 trained models with different architectures and training mixtures.	✓	0.21
Early-fusion architectures exhibit stronger performance at lower parameter counts compared to late-fusion architectures.	✓	0.27
Early-fusion architectures are more efficient to train than late-fusion architectures.	✓	0.24
Early-fusion architectures are easier to deploy than late-fusion architectures.	✓	0.25
Incorporating Mixture of Experts (MoEs) allows models to learn modality-specific weights, significantly benefiting perfo	✓	0.29

References

- https://www.semanticscholar.org/paper/5c0a82558db7a885bf6b174c2ed4e52de558a000
- https://www.semanticscholar.org/paper/0f7c74d0d990126bd96d8881dcca65286183ae35
- https://www.semanticscholar.org/paper/ecf10b245b492ad3033570500fdb504ecbc03a90