# SOVEREIGN: How does the optimal number of modality-specific experts in SMoES scale with VLM backbone size (e.g., 7B vs 13

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

**Abstract**

There has been a rapid progress in the task of Visual Question Answering with improved model architectures. Unfortunately, these models are usually computationally intensive due to their sheer size which poses a serious challenge for deployment. We aim to tackle this issue for the specific task of Visual Question Answering (VQA). A Convolutional Neural Network (CNN) is an integral part of the visual processing pipeline of a VQA model (assuming the CNN is trained along with entire VQA model). In this project, we propose an efficient and modular neural architecture for the VQA task with focus on

## 1 Introduction

Analysis of: Learning Sparse Mixture of Experts for Visual Question Answering. Research goal: How does the optimal number of modality-specific experts in SMoES scale with VLM backbone size (e.g., 7B vs 13B vs 34B) for ChartQA accuracy and throughput under distribution shift to unseen chart types?.

## 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. 6 claims extracted, 0 verified. Tribunal: 4.8/10 → 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 |
|---|---|---|
| The modular ResNeXt-32k (0% sparsity) achieves a baseline accuracy of 54.51% on the VQA v2 dataset. | × | 0.05 |
| At 50% sparsity, the model achieves 54.47% accuracy on VQA v2 dataset. | × | 0.07 |
| At 50% sparsity, there is a marked 3.62% loss in overall accuracy on the VQA v2 dataset. | × | 0.03 |
| The model with 75% sparsity has a 3.62% loss in overall accuracy on VQA v2 dataset. | × | 0.05 |
| The model with 0% sparsity has comparable performance with the one which doesn't have sparsity in the convolutional ResN | × | 0.07 |
| The model with 50% sparsity has comparable performance with the one which doesn't have sparsity in the convolutional Res | × | 0.07 |

## References


- http://arxiv.org/abs/1909.09192v1
- http://arxiv.org/abs/2504.13275v4
- http://arxiv.org/abs/2408.04852v1