Sparse Multimodal Model Efficiency and Alignment Trade-offs on VQAv2 and OK-VQA

Assignee Research

May 30, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: Does the inference efficiency of sparse multimodal models with varying numbers of experts improve with higher alignment scores on VQAv2 and OK-VQA, and how does this trade-off compare to dense models. Sparse Mixture-of-Experts (MoE) models can outperform dense large language models at similar computation by activating only a small set of experts per token. However, stacking many expert modules introduces substantial parameter memory, which makes MoE models difficult to deploy. 0 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 2.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: ExpertFlow: Efficient Mixture-of-Experts Inference via Predictive Expert Caching and Token Scheduling. Research question: Does the inference efficiency of sparse multimodal models with varying numbers of experts improve with higher alignment scores on VQAv2 and OK-VQA, and how does this trade-off compare to dense models?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 2.5/10.

3 Results

13 papers retrieved. 0 claims extracted; 0 independently verified. Quality review score: 2.5/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

References

- http://arxiv.org/abs/2505.11415v2
- http://arxiv.org/abs/2410.17954v2
- http://arxiv.org/abs/2402.14800v2