SOVEREIGN: How does SMoES-trained modality routing for multimodal LLMs generalize to out-of-distribution benchmarks like

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 2026

Abstract

Large-scale pre-trained models (PTMs) show great zero-shot capabilities. In this paper, we study how to leverage them for zero-shot visual question answering (VQA). Our approach is motivated by a few observations. First, VQA questions often require multiple steps of reasoning, which is still a capability that most PTMs lack. Second, different steps in VQA reasoning chains require different skills such as object detection and relational reasoning, but a single PTM may not possess all these skills. Third, recent work on zero-shot VQA does not explicitly consider multi-step reasoning chains, whic

1 Introduction

Analysis of: Modularized Zero-shot VQA with Pre-trained Models. Research goal: How does SMoES-trained modality routing for multimodal LLMs generalize to out-of-distribution benchmarks like DocVQA and InfographicVQA under domain shift, and what are the accuracy vs. latency trade-offs compared to chart-specific distribution shifts?.

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. 9 claims extracted, 1 verified. Tribunal: 4.7/10 → REJECT (revision_round=0). 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 GQA dataset consists of questions requiring multi-step reasoning and various reasoning skills.	×	0.11
Around 94% of the questions in the GQA dataset require multiple reasoning steps.	×	0.13
We regard GQA as the main dataset to demonstrate the effectiveness of the proposed method compared with the baselines.	×	0.10
Questions on the VQAv2 dataset require fewer reasoning steps and are of diverse semantics.	×	0.09
We evaluate the proposed modularized zero-shot VQA method on two benchmarks: GQA and VQAv2.	✓	0.20
We report standard accuracy for the GQA dataset while soft accuracy for VQAv2 dataset as there are multiple ground-truth	×	0.02
The proposed ModZero-VQA method is more effective on the GQA dataset, which contains many multi-step reasoning questions	×	0.09
Mod-Zero-VQA clearly surpasses CLIP.	×	0.05
Our ModZero-VQA method assigns each sub reasoning task to a pre-trained model capable of the task (i.e., MDETR for refer	×	0.12

References

- http://arxiv.org/abs/2305.17369v2
- http://arxiv.org/abs/1909.09192v1

- http://arxiv.org/abs/2507.22398v3