SOVEREIGN: Dynamic Clue Bottlenecks: Towards Interpretable-by-Design Visual Question Answer

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Recent advances in multimodal large language models (LLMs) have shown extreme effectiveness in visual question answering (VQA). However, the design nature of these end-to-end models prevents them from being interpretable to humans, undermining trust and applicability in critical domains. While post-hoc rationales offer certain insight into understanding model behavior, these explanations are not guaranteed to be faithful to the model. In this paper, we address these shortcomings by introducing an interpretable by design model that factors model decisions into intermediate human-legible explana

1 Introduction

Analysis of: Dynamic Clue Bottlenecks: Towards Interpretable-by-Design Visual Question Answering. Research goal: Does SMoES-style dynamic modality routing improve out-of-distribution robustness on VCR adversarial splits compared to Top-2 modality-agnostic MoE-VLMs with matched expert counts?.

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. 5 claims extracted, 5 verified. Tribunal: 7.8/10 → 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
DCLUB provides an explainable intermediate space before the VQA decision and is faithful from the beginning.	✓	0.26
DCLUB first returns a set of visual clues: natural language statements of visually salient evidence from the image, and	✓	0.33
A dataset of 1.7k reasoning-focused questions with visual clues was collected to supervise and evaluate the generation o	✓	0.30
DCLUB improves 4.64% over a comparable black-box system in reasoning-focused questions.	✓	0.22
DCLUB preserves 99.43% of performance of a comparable black-box system.	✓	0.15

References

- http://arxiv.org/abs/2305.14882v2
- http://arxiv.org/abs/2507.22398v3
- http://arxiv.org/abs/2406.06462v4