SOVEREIGN: Can Dynamic Clue Bottlenecks generalize to out-of-distribution robustness on VCR adversarial splits when scale

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 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: Can Dynamic Clue Bottlenecks generalize to out-of-distribution robustness on VCR adversarial splits when scaled to larger backbone LLMs (e.g., 7B vs 13B parameters), measured by accuracy drop under distribution shift?.

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. 10 claims extracted, 0 verified. Tribunal: 2.7/10 → 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
DCLUB achieves comparable VQA performance to blackbox models, with 74.8% accuracy on the DCLUB dataset compared to 71.48	×	0.06
On the VQA v2 dataset, DCLUB achieves 57.78% accuracy while the blackbox baseline achieves 58.11%.	×	0.03
On the GQA dataset, DCLUB achieves 40.00% accuracy while the blackbox baseline achieves 42.00%.	×	0.02
DCLUB's performance is 104.64% of the black-box baseline on the DCLUB dataset.	×	0.05
DCLUB's performance is 99.43% of the blackbox baseline on the VQA v2 dataset.	×	0.09
DCLUB's performance is 95.24% of the blackbox baseline on the GQA dataset.	×	0.03
The visual clue generator is fine-tuned on a base pre-trained BLIP-2 model using the 'pretrain flant5xl' checkpoint.	×	0.04
Training uses the image encoder ViT-L/14 from CLIP and frozen LLM FlanT5-XL (3B).	×	0.02
Training is completed using NVIDIA RTX A6000 (48G) GPUs.	×	0.02
The training data size is approximately 1.1k examples.	×	0.04

References

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

- http://arxiv.org/abs/2305.14882v2
- http://arxiv.org/abs/2103.15670v3