SOVEREIGN: Multi-Object Hallucination in Vision-Language Models

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 27, 2026

Abstract

Large vision language models (LVLMs) often suffer from object hallucination, producing objects not present in the given images. While current benchmarks for object hallucination primarily concentrate on the presence of a single object class rather than individual entities, this work systematically investigates multi-object hallucination, examining how models misperceive (e.g., invent nonexistent objects or become distracted) when tasked with focusing on multiple objects simultaneously. We introduce Recognition-based Object Probing Evaluation (ROPE), an automated evaluation protocol that consid

1 Introduction

Analysis of: Multi-Object Hallucination in Vision-Language Models. Research goal: How does the predictive expert caching strategy in ExpertFlow affect object existence hallucination rates (e.g., on POPE) compared to dense models of equivalent FLOPs in multimodal MoE-VLMs?.

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 $\rightarrow$ 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
Large vision language models (LVLMs) often suffer from object hallucination, producing objects not present in the given	✓	0.37
Current benchmarks for object hallucination primarily concentrate on the presence of a single object class rather than i	✓	0.34
LVLMs suffer more hallucinations when focusing on multiple objects compared to a single object.	✓	0.35
The tested object class distribution affects hallucination behaviors, indicating that LVLMs may follow shortcuts and spu	✓	0.35
Hallucinatory behaviors are influenced by data-specific factors, salience and frequency, and model intrinsic behaviors.	✓	0.28

References

- http://arxiv.org/abs/2407.06192v2
- http://arxiv.org/abs/2410.17954v2
- http://arxiv.org/abs/2604.23996v1