SOVEREIGN: Does the predictive expert caching strategy in ExpertFlow reduce object existence hallucination (POPE accuracy

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 28, 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: Does the predictive expert caching strategy in ExpertFlow reduce object existence hallucination (POPE accuracy) more effectively for rare vs. frequent object classes in multimodal MoE-VLMs, and how does this trade-off scale with model size and number of cached experts?.

2 Methodology

Multi-query arXiv search (1 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

3 papers retrieved. 5 claims extracted, 3 verified. Tribunal: 6.3/10 → REVISE (revision_round=1). 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
Vision-Language Models sometimes generate objects that do not exist in the provided images, which is referred to as obje	✓	0.16
ROPE is designed for evaluating multi-object hallucination and considers the distribution of object classes within a sin	✓	0.21
ROPE uses visual cues such as marked bounding boxes to refer to objects instead of textual descriptions.	×	0.06
ROPE evaluation is automated and does not require black-box neural models or human evaluators.	×	0.04
ROPE measures object hallucination in Vision-Language Models within a multi-object setting.	✓	0.22

References

- http://arxiv.org/abs/2412.06830v1
- http://arxiv.org/abs/2601.16325v1
- http://arxiv.org/abs/2407.06192v2