SOVEREIGN: How does the performance of multimodal models on Visual Genome benchmark tasks vary when trained with differen

SOVEREIGN Research Kernel
Autonomous draft — Owner review required before publication

May 29, 2026

Abstract

Despite progress in perceptual tasks such as image classification, computers still perform poorly on cognitive tasks such as image description and question answering. Cognition is core to tasks that involve not just recognizing, but reasoning about our visual world. However, models used to tackle the rich content in images for cognitive tasks are still being trained using the same datasets designed for perceptual tasks. To achieve success at cognitive tasks, models need to understand the interactions and relationships between objects in an image. When asked "What vehicle is the person riding?"

1 Introduction

Analysis of: Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. Research goal: How does the performance of multimodal models on Visual Genome benchmark tasks vary when trained with different vision-language pretraining objectives, measured by caption generation BLEU scores and visual question answering accuracy across object, attribute, and relationship prediction subtasks?.

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

10 papers retrieved. 7 claims extracted, 7 verified. Tribunal: 7.5/10 $\rightarrow$ 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
The Visual Genome dataset contains over 108K images.	✓	0.17
Each image in the Visual Genome dataset has an average of 35 objects, 26 attributes, and 21 pairwise relationships.	✓	0.25
The dataset contains over 108K images where each image has an average of 35 objects.	✓	0.21
Visual Genome dataset contains objects, attributes, and relationships annotations.	✓	0.23
The model needs to understand the interactions and relationships between objects in an image to answer questions such as	✓	0.24
The dataset includes canonicalized objects, attributes, relationships, and noun phrases in region descriptions and quest	✓	0.28
Computers perform poorly on cognitive tasks such as image description and question answering.	✓	0.28

References

- https://openalex.org/W3167118264
- https://doi.org/10.1109/cvpr.2019.00207
- https://doi.org/10.1007/s11263-016-0981-7