Multimodal Reasoning Accuracy in Generative AI Models: Benchmarking DALL-E 2 and Stable Diffusion

Assignee Research

June 2, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How do different generative AI models (e.g., DALL-E 2, Stable Diffusion) perform in terms of multimodal reasoning accuracy when evaluated on benchmarks like MiniGPT-4 and LLaVA, and how does. 16 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.4/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: LLaVA-LE: Large Language-and-Vision Assistant for Lunar Exploration. Research question: How do different generative AI models (e.g., DALL-E 2, Stable Diffusion) perform in terms of multimodal reasoning accuracy when evaluated on benchmarks like MiniGPT-4 and LLaVA, and how does alignment training influence their performance?.

2 Methodology

Systematic literature search across multiple databases yielded 15 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.4/10.

3 Results

15 papers retrieved. 16 claims extracted; 1 independently verified. Quality review score: 4.4/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
The evaluation benchmark consists of 50 lunar patches from the LUCID stage 1 dataset.	×	0.06
190 questions were generated for the evaluation benchmark using a language-only GPT-5.1 model.	×	0.05
The evaluation set was excluded from both stages of training LLaVA-LE.	×	0.05
Base LLaVA is the original LLaVA-v1.5-13B model trained on general-purpose multimodal instruction data without lunar-spe	×	0.09
Concept-Aligned LLaVA-LE (Stage 1) is trained on caption-based concept alignment data without instruction tuning.	×	0.12
Fully Trained LLaVA-LE (Stage 2) incorporates both Stage 1 alignment and Stage 2 instruction tuning on the LUCID VQA dat	×	0.09
LLaVA-LE Stage 2 achieves an average overall score of 0.921 when averaged across GPT and Gemini judges.	×	0.12
LLaVA-LE Stage 2 represents a 3.3$\times$ improvement over Base LLaVA, which scored 0.278.	×	0.08
LLaVA-LE Stage 2 represents a 2.1$\times$ improvement over LLaVA-LE Stage 1, which scored 0.443.	×	0.07
LLaVA-LE Stage 2 scored 1.070 on the Reasoning category, exceeding the judge's own reference score.	✓	0.16
LLaVA-LE Stage 2 scored 0.922 on the Detailed category.	×	0.07
LLaVA-LE Stage 2 scored 0.698 on the Conversation category.	×	0.05
The LUCID dataset contains 96K samples derived from co-registered lunar remote sensing observations.	×	0.08
Stage 1 of the LUCID dataset contains 76K samples used for concept alignment.	×	0.07
Stage 2 of the LUCID dataset contains approximately 20K samples used for instruction tuning.	×	0.11
There is currently no publicly available multimodal dataset in planetary science that supports the training of vision-la	×	0.13

References

- http://arxiv.org/abs/2603.24696v1
- http://arxiv.org/abs/2503.14504v2
- http://arxiv.org/abs/2408.07303v2