Visual Complexity Effects on Reasoning Accuracy in LLaVA-NeXT and Video-LLaVA-8B

Assignee Research

June 6, 2026

Abstract

This report synthesises findings from 13 peer-reviewed papers addressing the following research question: What is the impact of varying levels of visual complexity in diagrams on the reasoning accuracy of LLaVA-NeXT and Video-LLaVA-8B, and how does this correlate with their performance on standard. 14 claims were extracted from source literature; 2 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Visual Reasoning Tracer: Object-Level Grounded Reasoning Benchmark. Research question: What is the impact of varying levels of visual complexity in diagrams on the reasoning accuracy of LLaVA-NeXT and Video-LLaVA-8B, and how does this correlate with their performance on standard vision-language benchmarks?.

2 Methodology

Systematic literature search across multiple databases yielded 13 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.5/10.

3 Results

13 papers retrieved. 14 claims extracted; 2 independently verified. Quality review score: 4.5/10.

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
Many advanced reasoning models produce correct final textual answers but fail to generate intermediate visual reasoning	✓	0.16
Current models' internal grounding remains opaque even when the final output is correct.	×	0.10
Applying supervised fine-tuning (SFT) and Reinforcement Learning (RL) on the VRT-80k dataset enables an MLLM with pixel-	×	0.07
The Visual Reasoning Tracer (VRT) is the first study to explore interactive, joint outputs of reasoning text and visual	×	0.12
The authors propose a novel metric called Visual Quality (VQ) to jointly assess the quality of the reasoning process and	×	0.07
The VRT-80k dataset is constructed as a large-scale, high-quality training dataset for the visual reasoning trace task.	✓	0.15
Evaluation on VRT-Bench provides empirical evidence that current SOTA MLLMs fail to ground their intermediate reasoning	×	0.13
The R-Sa2VA-Qwen3VL-4B-SFT model achieves an Overall R-LQ score of 66.3, an R-VQ score of 87.3, and an Accuracy (A) of 5	×	0.04
Adding Reinforcement Learning (RL) to the R-Sa2VA-Qwen3VL-4B-SFT model improves the Overall Accuracy (A) from 59.5 to 62	×	0.02
Adding both RL and Segmentation Loss to the base model results in an Overall Accuracy (A) of 61.1.	×	0.03
The Q3-4B-Joint model achieves scores of 79.1 on RefCOCO, 74.5 on RefCOCO+, and 77.4 on RefCOCOg.	×	0.01
Removing the filter component from the R-Sa2VA-Qwen3VL-4B-RL model results in an Overall Accuracy (A) of 61.8.	×	0.02
Previous visual reasoning works implement reasoning processes that are purely linguistic and not sufficiently grounded i	×	0.08
The VRT task requires models to align each reasoning step with a visual segmentation mask.	×	0.12

References

- http://arxiv.org/abs/2512.05091v1
- http://arxiv.org/abs/2508.17298v2
- http://arxiv.org/abs/2209.10326v2