Multi-Turn Conversational Impact on VLM Accuracy and Coherence in MULTIVERSE

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 9 peer-reviewed papers addressing the following research question: What is the impact of increasing the number of conversational turns in MULTIVERSE on the accuracy and coherence of responses from state-of-the-art VLMs. Vision-and-Language Models (VLMs) have shown impressive capabilities on single-turn benchmarks, yet real-world applications often demand more intricate multi-turn dialogues. Existing multi-turn datasets (e.g, MMDU, ConvBench) only partially capture the breadth and depth of. 16 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: MultiVerse: A Multi-Turn Conversation Benchmark for Evaluating Large Vision and Language Models. Research question: What is the impact of increasing the number of conversational turns in MULTIVERSE on the accuracy and coherence of responses from state-of-the-art VLMs?.

2 Methodology

Systematic literature search across multiple databases yielded 9 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 3.2/10.

3 Results

9 papers retrieved. 16 claims extracted; 0 independently verified. Quality review score: 3.2/10.

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
VPGTrans's high precision (88) with its lower recall (65) under adversarial conditions	×	0.01
49,700 images were obtained from the test sets of MMMU [73], MMMU-Pro [74], NaturalBench [37], VisIT-Bench [6], MathVist	×	0.01
13,998 images (29.1%) were discarded due to deduplication using pHash	×	0.01
21,808 images (64.77%) were removed due to scoring below 5 in clarity, resolution, and real-world plausibility	×	0.03
213 images (1.80%) were excluded due to having fewer than 50 images in their category	×	0.04
The most common image categories were Charts and Graphs (22.92%) and Diagrams and Schematics (11.17%)	×	0.04
The total number of selected images was capped at 1K	×	0.00
Personal backgrounds were generated to create more plausible and realistic multi-turn conversations	×	0.10
InternVL2.5-1B achieved scores of 13.93, 21.36, 23.51, and 26.08 in Turns 1 to 4 respectively	×	0.01
Qwen2.5-VL-72B achieved scores of 5, 5, 5, and 5 in Turns 1 to 4 respectively	×	0.01
Larger models generally perform better across tasks, but scaling effects vary	×	0.04
Qwen2.5-VL-72B excels in structured reasoning tasks (e.g., mathematics, coding)	×	0.09
Qwen2.5-VL-7B shows stronger creative abilities	×	0.01
R2=0.08, p=6.82e-14 for Turn 1	×	0.03
R2=0.01, p=1.16e-02 for Turn 2	×	0.03
R2=0.03, p=7.42e-05 for Turn 3	×	0.03

References

- http://arxiv.org/abs/2210.02526v1
- http://arxiv.org/abs/2510.16641v1
- http://arxiv.org/abs/2103.08733v1