MobileVLM Accuracy on MME and MM1K Against Quantized 3B–13B VLMs

Assignee Research

June 5, 2026

Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How does the accuracy of MobileVLM's 1.4B and 2.7B models on the MME and MM1K benchmarks compare to quantized versions of larger 3B to 13B VLMs. 12 claims were extracted from source literature; 12 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 8.5/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: RoboMamba: Efficient Vision-Language-Action Model for Robotic Reasoning and Manipulation. Research question: How does the accuracy of MobileVLM's 1.4B and 2.7B models on the MME and MM1K benchmarks compare to quantized versions of larger 3B to 13B VLMs?.

2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 8.5/10.

3 Results

4 papers retrieved. 12 claims extracted; 12 independently verified. Quality review score: 8.5/10.

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
A fundamental objective in robot manipulation is to enable models to comprehend visual scenes and execute actions.	✓	0.28
Existing Vision-Language-Action (VLA) models for robots can handle a range of basic tasks.	✓	0.31
Existing VLA models face challenges in insufficient reasoning ability to tackle complex tasks.	✓	0.22
Existing VLA models face challenges in high computational costs for fine-tuning and inference.	✓	0.23
The state space model (SSM) known as Mamba demonstrates promising capabilities in non-trivial sequence modeling with lin	✓	0.34
RoboMamba is an end-to-end robotic VLA model that leverages Mamba to deliver both robotic reasoning and action capabilit	✓	0.44
RoboMamba integrates the vision encoder with Mamba, aligning visual tokens with language embedding through co-training.	✓	0.24
RoboMamba is empowered with visual common sense and robotic-related reasoning.	✓	0.19
RoboMamba explores an efficient fine-tuning strategy with a simple policy head to equip it with SE(3) pose prediction ab	✓	0.28
RoboMamba can acquire manipulation skills with minimal fine-tuning parameters (0.1% of the model) and time once it posse	✓	0.31
RoboMamba demonstrates outstanding reasoning capabilities on general and robotic evaluation benchmarks.	✓	0.27
RoboMamba showcases impressive pose prediction results in both simulation and real-world scenarios.	✓	0.20

References

- https://doi.org/10.48550/arxiv.2405.10739
- https://doi.org/10.48550/arxiv.2306.13549
- https://doi.org/10.48550/arxiv.2406.04339