Validating SafeCoDe Model-Agnosticism Across Multimodal Architectures and Assessing Cross-Domain Robustness

Assignee Research

June 13, 2026

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated exceptional performance in artificial intelligence by facilitating integrated understanding across diverse modalities, including text, images, video, audio, and speech. However, their deployment in real-world applications raises significant concerns about adversarial vulnerabilities that could compromise their safety and reliability. Unlike unimodal models, MLLMs face unique challenges due to the interdependencies among modalities, making them susceptible to modality-specific threats and cross-modal adversarial manipulations. This pape

1 Introduction

This paper examines: Survey of Adversarial Robustness in Multimodal Large Language Models. Research question: Can the model-agnostic nature of SafeCoDe be validated across different multimodal architectures (e.g., LLaVA, Qwen-VL) by comparing their safety alignment scores on benchmarks like MMBench, and what is the impact on cross-domain robustness?.

2 Methodology

Systematic literature search across multiple databases yielded 18 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 7.6/10.

3 Results

18 papers retrieved. 15 claims extracted; 13 independently verified. Quality review score: 7.6/10.

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
Chowdhury et al. (2024) focus on text-based attacks in LLMs.	✓	0.27
Liu et al. (2024b) provide a macroscopic review of LVLM safety issues limited to vision and text modalities.	✓	0.24
Liu et al. (2024a) analyze vision-text attacks but neglect other modalities.	✓	0.18
Fan et al. (2024) narrow their analyses to image-based attacks.	✓	0.25
MLLMs integrate multiple modalities into a unified framework enabling cross-modal reasoning for tasks like captioning, r	✓	0.19
MLLMs leverage large language model backbones, such as GPT-based models, to process textual data.	✓	0.22
Specialized encoders in MLLMs extract features from non-textual modalities.	×	0.12
MLLMs rely on fusion modules that align and combine multimodal features, often mapping them into shared latent spaces.	✓	0.23
LlamaPartialSpoof is used to evaluate robustness against spoofing attacks in speech-based applications.	✓	0.15
ActivityNet-200 and MSVD-QA are datasets in the video domain focusing on evaluating models' ability to capture temporal	✓	0.23
LibriSpeech is a dataset used for audio-based MLLMs.	×	0.11
Adversarial examples generated using diffusion models can effectively transfer across various vision-language models, in	✓	0.20
AdvDiffVLM enhances transfer-based attacks by employing ensemble gradient estimation and mask generation techniques.	✓	0.24
The Unicorn safety evaluation framework provides a systematic evaluation of MLLM vulnerabilities under out-of-distributi	✓	0.20
Findings from the Unicorn benchmark suggest that visual language training pipelines often weaken original safety measure	✓	0.20

References

- https://arxiv.org/abs/2503.13962
- https://arxiv.org/abs/2508.15370
- http://arxiv.org/abs/2509.19212v1