Multimodal Soft Prompt Attacks and Alignment Robustness in Open-Source Models

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 15 peer-reviewed papers addressing the following research question: How do multimodal soft prompt attacks (e.g., combining text and image embeddings) affect the robustness of alignment in open-source multimodal models like LLaVA, compared to text-only attacks. Although multimodal large language models (MLLMs) are increasingly deployed in real-world applications, their instruction-following behavior leaves them vulnerable to prompt injection attacks. Existing prompt injection methods predominantly rely on textual prompts or perceptible. 15 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 3.0/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Adversarial Prompt Injection Attack on Multimodal Large Language Models. Research question: How do multimodal soft prompt attacks (e.g., combining text and image embeddings) affect the robustness of alignment in open-source multimodal models like LLaVA, compared to text-only attacks?.

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 3.0/10.

3 Results

15 papers retrieved. 15 claims extracted; 0 independently verified. Quality review score: 3.0/10.

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
CoTTA is a novel attack framework that induces specified malicious sentences from closed-source MLLMs via imperceptible	×	0.14
CoTTA proposes a covert text trigger as an additional textual noise alongside adversarial perturbations.	×	0.05
CoTTA designs an adaptive target image that is iteratively updated to bridge cross-modal representations.	×	0.07
Experiments were conducted on two tasks against various powerful closed-source MLLMs including GPT-4o, GPT-5, Gemini-2.5	×	0.13
On the GPT-4o model, the CoTTA method (ours) achieved an Attack Success Rate (ASR) of 81% and an AvgSim of 0.442.	×	0.06
On the GPT-4o model, the CoTTA method outperformed FOA-Attack, which achieved an ASR of 47% and an AvgSim of 0.232.	×	0.02
On the GPT-5 model, the CoTTA method achieved an ASR of 56% and an AvgSim of 0.320.	×	0.02
On the Gemini-2.5 model, the CoTTA method achieved an ASR of 79% and an AvgSim of 0.429.	×	0.02
On the Claude-4.5 model, the CoTTA method achieved an ASR of 8% and an AvgSim of 0.046.	×	0.02
Removing the covert trigger component from CoTTA reduced the ASR from 81% to 66% and AvgSim from 0.442 to 0.306 on GPT-4	×	0.01
Removing the image-to-text alignment component from CoTTA resulted in an ASR of 75% and AvgSim of 0.42 on GPT-4o.	×	0.06
Removing the image-to-image alignment component from CoTTa reduced the ASR significantly to 27% and AvgSim to 0.155 on G	×	0.03
Disabling the updating of the target image in CoTTA resulted in an ASR of 73% and AvgSim of 0.434 on GPT-4o.	×	0.02
AttackVLM, AnyAttack, M-Attack, FOA-Attack, and Agent-Attack all achieved 0% ASR on GPT-4o when using Laion or specific	×	0.03
CoTTA consistently outperforms existing approaches by a large margin in extensive experiments.	×	0.05

References

- http://arxiv.org/abs/2603.29418v1
- http://arxiv.org/abs/2312.03777v2
- http://arxiv.org/abs/2405.18770v6