Responsible NLP Checklist

Paper title: *Reference Attack: A New Cross-Modal Jailbreaking Attack against Multimodal Large Language Models*
Authors: *Yulong Wang, Yifei Fu, Jiayi Gao*

How to read the checklist symbols:

☑ the authors responded 'yes'

☒ the authors responded 'no'

N/A the authors indicated that the question does not apply to their work

☐ the authors did not respond to the checkbox question

For background on the checklist and guidance provided to the authors, see the Responsible NLP Checklist page at ACL Rolling Review.

☑ **A. Questions mandatory for all submissions.**

☑ A1. Did you describe the limitations of your work?
This paper has a Limitations section.

☑ A2. Did you discuss any potential risks of your work?
Section 3.5, Appendix G, Appendix H, and Ethical Considerations section.

☑ **B. Did you use or create scientific artifacts? (e.g. code, datasets, models)**

☑ B4. Did you discuss the steps taken to check whether the data that was collected/used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect/anonymize it?
Section 3.1 and Appendix D.2 describe the datasets used. SafeBench contains 500 malicious text queries and MM-SafetyBench contains 5,040 text-image pairs across safety-critical scenarios. These are publicly available benchmark datasets specifically designed for safety evaluation. The datasets contain intentionally harmful prompts for research purposes but do not contain personally identifying information. All experiments follow responsible disclosure practices as described in the Ethical Considerations section.

☑ B6. Did you report relevant statistics like the number of examples, details of train/test/dev splits, etc. for the data that you used/created?
Section 3.1 and Appendix D.2 provide dataset statistics including number of examples (500 for SafeBench, 5,040 for MM-SafetyBench) and the 13 safety-critical scenarios covered.

☑ **C. Did you run computational experiments?**

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3.1 describes experimental setup including metric (ASR), evaluation protocol, datasets, models tested, and baselines. Appendix D provides detailed settings including LLM-as-judge protocol, human validation, and attack execution. For LLaMA 3.2, hyperparameters are specified in Appendix C.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean,

The Responsible NLP Checklist used at ACL Rolling Review is adopted from NAACL 2022, with the addition of ACL 2023 question on AI writing assistance and further refinements based on ARR practice. ACL 2026 used a subset of ARR checklist form.

etc. or just a single run?
Tables 1, 2, and 3 report attack success rates (ASR) across models and categories. Section 3.1 describes evaluation using majority voting across three critic LLMs with human validation on 100 outputs per method. Each attack was performed three times (Appendix D.4); success required at least one successful attempt.

☑ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix D.1 describes that human annotators assessed whether responses violated AI safety policies without knowledge of the generating model or attack technique. Response ordering was randomized to eliminate biases. The evaluation task was binary (policy violation: yes/no).

☒ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
While human validation was conducted (Appendix D.1), recruitment and payment details were not included as the validation was limited in scope (100 samples per method per model) and served only to validate the automated evaluation protocol rather than being the primary evaluation method.

N/A D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating (e.g., did your instructions explain how the data would be used)?
The paper uses publicly available benchmark datasets (SafeBench and MM-SafetyBench) specifically designed for safety evaluation research. Human annotators evaluated model outputs, not human-generated data requiring consent.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
This research evaluates security vulnerabilities of publicly available AI systems using publicly available benchmark datasets. No human subjects research requiring IRB approval was conducted. Human validation involved only evaluation of AI-generated outputs, not collection of personal data or behavioral studies. The research follows responsible disclosure practices (Ethical Considerations section).

☑ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

☑ E1. If you used AI assistants, did you include information about their use?
We used AI assistants for automated evaluation of attack success rates. Specifically, three large language models (ChatGPT-4o, Claude 3.5 Sonnet, and Gemini 2.0 Flash) served as critic judges to assess whether target model responses violated AI safety policies. Each response was independently evaluated by all three critics, with final success determined by majority voting (Section 3.1). The prompt given to critic models is provided in Appendix D.1. No AI assistants were used in the research design, attack development, data collection, or manuscript writing.