Responsible NLP Checklist

Paper title: *VisCRA: A Visual Chain Reasoning Attack for Jailbreaking Multimodal Large Language Models*
Authors: *Bingrui Sima, Linhua Cong, Wenxuan Wang, Kun He*

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?
We discuss this in the Ethical Statement on page 9

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

☑ B1. Did you cite the creators of artifacts you used?
Section 5.1 (Experimental Setup) explicitly names and cites the creators of all models and benchmarks used in our study, such as HADES, MM-SafetyBench, and the various open- and closed-source MLLMs.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We confirm that our use of all artifacts adheres to their respective licenses in our Ethical Statement (Page 9).

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
We state that our use of existing artifacts is consistent with their intended use in our Ethical Statement (Page 9).

☑ 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?
Our work uses safety benchmarks like HADES and MM-SafetyBench, which are explicitly designed to contain and represent harmful/offensive content categories for the purpose of safety evaluation. We acknowledge this with a content warning on page 1. These are established public benchmarks, and we assume they have been curated to exclude personally identifying information (PII).

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.

☒ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We use existing, publicly available, and well-documented artifacts (models and benchmarks). We do not create or release new artifacts that would require additional documentation from our side.

☑ 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 5.1, under the "Baselines and Benchmarks" paragraph, provides statistics for the datasets used, including the number of samples for HADES (750) and the subset of MM-SafetyBench (741).

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

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
We report model sizes for some models and total computational budget in Section 5.1 .

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We discuss the experimental setup and key hyperparameters for our method in Section 5.1 under Implementation Details. Furthermore, we provide a detailed hyperparameter ablation study for our masking component in Appendix A.2.

☑ 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, etc. or just a single run?
We report the Attack Success Rate (ASR), which is the mean success rate over all samples in the respective benchmarks. This makes it transparent that we are reporting a summary statistic over a full set of experiments, not the result of a single run. The methodology is described in Section 5.1 under Evaluation Metrics.

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
Our method does not rely on standard NLP packages (e.g., NLTK, SpaCy) for preprocessing or analysis. Our evaluation is based on an LLM evaluator (Llama-Guard-3-8B), as specified in Section 5.1.

☒ **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.?
(left blank)

☒ 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)?
(left blank)

☒ 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)?
(left blank)

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
(left blank)

☒ N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
(left blank)

☑ **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?
Ethical Statement (Page 9).