Responsible NLP Checklist

Paper title: *Mitigating Safety Context Amnesia in Multimodal Reasoning Models via Intent-Guided Safety Reasoning*
Authors: *Xiyao Dong, Guangsheng Cheng, YiLong Chen, Xiaojin Zhang, 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?
Section 5.2 and 5.3

☑ **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?
We utilized standard, publicly available datasets (VLGuard, VLSbench, MM-SafetyBench, etc.) that are widely established in the field. We did not collect new data from human subjects, so we relied on the anonymization protocols of the original dataset creators. Regarding offensive content, these datasets explicitly contain harmful examples (e.g., hate speech, violence) which are necessary and intentional for evaluating the safety defense mechanisms proposed in this work.

☑ 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 (Experimental Setup) and Appendix A (Dataset Composition) provide detailed statistics on the training data size (3,881 samples), split between safety and helpfulness categories, and the specific composition of evaluation benchmarks like VLSbench and MM-SafetyBench.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.1 outlines the experimental setup. Appendix A.1 and Table 5 explicitly list the hyperparameters used for fine-tuning the Perception Decoupler, including LoRA rank, learning rate, batch size, and scheduler details.

☑ 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?

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.

Section 5.2 and Tables 1-3 report the Defense Success Rate (DSR) and other metrics for all compared models. The results represent the performance on the standardized test sets.

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

N/A 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.?
No human subjects were involved.

N/A 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)?
No human subjects were involved.

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)?
No human subjects were involved.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No human subjects were involved.

☒ **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 utilized GPT-4o as a teacher model for Structured Evidence Synthesis (Section 4.2.1) and as an automated safety evaluator for calculating metrics (Section 5.1 and Appendix A.4).