Responsible NLP Checklist

Paper title: *Visual Self-Fulfilling Alignment: Shaping Safety-Oriented Personas via Threat-Related Images*
Authors: *Qishun Yang, Shu Yang, Lijie Hu, Di Wang*

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?
See Limitations section, paragraph 4.

☑ **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?
All images are AI-generated and contain no real individuals. VQA pairs are neutral questions about image content. No personally identifying information was collected.

☑ 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?
See Section 3 (Method).

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
See Section 3 (Method) and Section 4 (Experiments).

☒ 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 single-run results without error bars. This follows standard practice in VLM safety research. Prior work including VLGuard (ICML 2024) and AdaShield (ECCV 2024) also reports single-run results. Our experiments span 4 models 4 methods 3 benchmarks, and results show consistent trends across all conditions, suggesting stability.

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.

☒ **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.?
We did not use human annotators. Training data (images and VQA pairs) were generated by AI models. Evaluation was conducted using GPT-4o as the judge.

☒ 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 participants were recruited.

☒ 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-generated data requiring consent was collected.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No human subjects research was conducted.

☑ **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 GPT-5 for grammar checking and language polishing of the manuscript. All scientific content, experimental design, and analysis are the authors' original work.