Responsible NLP Checklist

Paper title: *ATAAT: Adaptive Threat-Aware Adversarial Tuning Framework against Backdoor Attacks on Vision-Language-Action Models*
Authors: *Kewei Chen, Yayu Long, Shuai Li, Mingsheng Shang*

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 the section 7 "Ethical Considerations". We explicitly acknowledge the dual-use risks of backdoor injection techniques and discuss mitigation through theoretical analysis and physical safety controls.

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

N/A 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?
See section 7 "Ethical Considerations". We explicitly state that the human subjects in the physical experiments are the authors themselves, and no personally identifiable information such as full faces was recorded or released.

☑ 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 4.1 "Experimental Setup" (paragraph "Datasets and Benchmarks") and Appendix D "Experimental Setup and Hyperparameters" (paragraph "Dataset and Poisoning Settings"), where we report the dataset tasks, poisoning rates (5%), and the number of few-shot anchor samples (200).

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
See Section 4.1 "Experimental Setup" and Appendix D "Experimental Setup and Hyperparameters". Specifically, Table 8 provides a comprehensive list of hyperparameters, including learning rate, batch size, LoRA rank, and perturbation budgets.

☑ 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?
See Section 4. We reported the mean Success Rate (SR) and Targeted Attack Success Rate (TASR) in

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.

Table 1. Additionally, in Figure 4, we explicitly visualized the standard deviation (represented by the shaded area) across multiple runs to demonstrate the stability of the gradient similarity metric.

☑ **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.?
The "human subjects" involved in the physical experiments were the authors themselves demonstrating the system. No external participants were recruited, so no instructions were needed.

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)?
The participants were the authors themselves, and no recruitment or payment was involved.

☒ 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)?
Since the subjects were the authors, consent was inherent. We ensured that no personally identifying information (such as faces) was captured in the visual data.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The study involved only the authors conducting technical validation of the proposed algorithms. No external human subjects or crowdworkers were recruited, and no sensitive personal data was collected. Therefore, IRB approval was not applicable.

☑ **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?
See "Appendix H: AI Use Declaration" at the end of the manuscript. We explicitly state that AI tools were used solely for grammatical error correction, sentence structure refinement, and notation consistency checking, while all conceptualization and experiments were performed by the authors.