Responsible NLP Checklist

Paper title: *Spatiotemporal Sycophancy: Negation-Based Gaslighting in Video Large Language Models*
Authors: *Ziyao Tang, Pengkun Jiao, Bin Zhu, Huiyan Qi, Jingjing Chen, Yu-Gang Jiang*

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?
Potential risks and the broader impact of identifying Vid-LLM vulnerabilities are discussed in the "Ethics Statement" on page 10. We emphasize that our goal is to enhance AI robustness rather than promote harmful exploitation.

☑ **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 address data privacy and quality in the "Quality Control" (Section 3.3) and "Data Usage Statement" on page 10. We utilized publicly available benchmarks where consent was handled by original creators and conducted manual reviews to filter out invalid or ambiguous content.

☑ 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?
Detailed statistics for the Gas Video-1000 benchmark, including the total sample count (1,013) and distribution across 10 categories, are provided in Section 3.3 and Figure 3.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The experimental environment, including the use of eight NVIDIA RTX 6000 GPUs and official APIs for model inference, is detailed in Section 7.9. Models were evaluated using default settings as described in Section 4.1.

☑ 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 descriptive metrics including the Sycophancy Rate (SR) and accuracy gaps (ΔAcc), specifying that results are derived from observable text outputs of single runs as detailed in Section 2.4 and Section 4.1.

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.?
Two annotators manually reviewed candidate questions as described in the "Quality Control" portion of Section 3.3.

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)?
As stated in the "Data Usage Statement" on page 10, no new data involving human subjects were collected.

☑ 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)?
Data Usage Statement. We use publicly available benchmark datasets released for research purposes, with consent handled by the original dataset creators.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
(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?
LLM Usage Statement. LLMs were used only for language polishing, grammar correction, and stylistic refinement during manuscript preparation.