Responsible NLP Checklist

Paper title: *CCTVBench: Contrastive Consistency Traffic VideoQA Benchmark for Multimodal LLMs*
Authors: *Xingcheng Zhou, Hao Guo, Rui Song, Walter Zimmer, Mingyu Liu, Andr Schamschurko, Hu Cao, Alois Knoll*

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?
Limitations; Section 4.5 (discussion of dataset bias, synthetic artifacts)

☑ **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?
No. The benchmark is built from publicly available traffic/anomaly videos and generated counterfactual counterparts, and the paper does not describe a dedicated screening or anonymization procedure for personally identifying or offensive 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?
Sections 3.2, 4.4, and Appendix A.2.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Sections 4.2 and Appendix A.3.3.

☑ 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?
Section 4.4 and Table 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.?
No. The paper describes the annotation and validation pipeline, but it does not provide the full text of annotator instructions and screenshots.

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.

☒ 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. The annotation and validation were conducted internally by the authors/research team, without external participant recruitment or monetary compensation.

☒ 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. The benchmark is built from publicly available traffic/anomaly videos and generated counterfactual counterparts.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No. The work uses publicly available traffic videos and internally curated annotations/validation, and does not involve a new human-subject data collection protocol requiring ethics review board approval.

☑ **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?
Section 3.1 and Appendix A.3. AI assistants were also used for language editing and clarity improvement, and for coding assistance such as code completion and debugging.