Responsible NLP Checklist

Paper title: *DRIVINGVQA: A Dataset for Interleaved Visual Chain-of-Thought in Real-World Driving Scenarios*
Authors: *Charles Corbire, Simon Roburin, Syrielle Montariol, Antoine Bosselut, Alexandre Alahi*

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?
The data do not include personally identifiable information, sensitive content, or user-generated material, and all experiments are conducted in a research-only context. Because the dataset and methods do not pose foreseeable ethical, safety, or societal risks, a dedicated discussion of potential risks was not included.

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

☑ B1. Did you cite the creators of artifacts you used?
All reused datasets and models (A-OKVQA, LLaVA-OV, GroundingDINO, GPT-4o) are properly cited (see References and Sections 2.24.2).

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
DrivingVQA is derived from publicly available driving theory exams (Section 3.1), and code/dataset will be made publicly available (Abstract, line 28). Licensing details follow dataset availability norms (open, educational, non-commercial use).

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Reused datasets and open source models (e.g., A-OKVQA, LLaVA-OV, GroundingDINO) are used for research and benchmarking, consistent with their intended academic purpose (Section 5).

☒ 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?
Discussed in Section 3.1 (Data Collection) images are from public driving exams, containing no personal data; all samples are reviewed and filtered for appropriateness using GPT-4o and manual verification.

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.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Comprehensive dataset statistics are presented in Figure 3 and Section 3.2, covering entity counts, label distributions, and annotation details.

☑ 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?
Reported in Section 3.2 and Figure 3 includes number of samples (3,931), entities (5,657), train/test splits, and entity distributions.

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

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4.2 specifies the model (LLaVA-OV-7B), training epochs (10), and fine-tuning setup (GPU infrastructure implied). Specific compute hours are not given but model scale and hardware are described.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Detailed in Section 4.2 and Appendix D.1, including fine-tuning epochs, seeds, optimizer setup, and architecture 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?
Table 1 reports mean standard deviation over 5 seeds; Section 4.2 describes evaluation metrics and variation.

N/A C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
(left blank)

☑ **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.?
Appendix B.1B.2 provides prompts and annotation instructions for human reviewers.

☑ 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)?
Human annotators were experts reviewing pseudo-annotations, not crowdworkers (Section 3.2). No monetary or demographic-sensitive recruitment; hence no payment concerns.

☑ 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)?
The annotators were explained the purpose of the study. Appendix B provides instructions they were given.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No personal nor sensitive data was collected, no ERB was needed.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Annotators are described as human experts in driving theory (Section 3.2, Appendix B.1).

☑ **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?
AI tools such as GPT-4o and GPT-4o-mini were used for translation, filtering, and generating interleaved reasoning traces (Sections 3.13.3, Appendix B.2).