Responsible NLP Checklist

Paper title: *GameplayQA: A Benchmarking Framework for Decision-Dense POV-Synced Multi-Video Understanding of 3D Virtual Agents*
Authors: *Yunzhe Wang, Runhui Xu, Kexin Zheng, Tianyi Zhang, Jayavibhav Niranjan Kogundi, Soham Hans, Volkan Ustun*

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 dataset is derived from non-sensitive simulated gameplay and does not involve personal data or real-world deployment, and therefore we did not identify specific potential risks beyond the general limitations discussed in the paper.

☑ **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?
The dataset consists of gameplay videos from simulated environments sourced from publicly available YouTube and Twitch videos, along with corresponding annotations. The data does not involve real human subjects, personally identifying information, or personal data, and therefore no anonymization or protection procedures were required.

☑ 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?
We report detailed statistics regarding number of questions/videos/duration of each type in table 2, as well as annotation statistics in section 3 and in Appendix C

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We use commercial MLLM API, where model setting are in table 8 and table 9 in appendix

☑ 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 3, table 4, table 5, table 6 with detail explanation in results write-up. We report results across different set-up with mean values for each model and over all models

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.?
in appendix E. Annotation Protocol

☑ 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)?
in section 3.2 Multi-Video Timeline Captioning; Section 3.3, human evaluation; and Appendix E.1

☑ 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)?
in appendix E.1. Annotators were informed of the purpose of the data collection, and They consented to their annotations being used for research and potential release

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The study involves annotators performing non-sensitive captioning of gameplay videos and does not collect personal or identifying information. Under our institutions guidelines, this type of data annotation task does not require ethics 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?
In section Acknowledgement. The authors acknowledge the use of Large Language Models for assistance with proofreading and grammar checking. All content was reviewed, edited, and approved by the human authors, who take full responsibility for the final manuscript.