Responsible NLP Checklist

Paper title: *EmoS: A HighFidelity Multimodal Benchmark for Fine-grained Streaming Emotional Understanding*
Authors: *Pengze Guo, Jingxi Liang, Zhiwen Xie, Qifeng Wang, Derek F. Wong*

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?
It does not involve direct deployment in high-stakes real-world environments or interactive human subject experiments, hence a dedicated discussion on potential risks was not included.

☑ **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 is constructed exclusively from publicly available broadcast media (e.g., TV shows, movies, and public platforms as detailed in Appendix A), featuring professional actors and public figures.Therefore, it does not contain private personally identifying information (PII) of unconsenting individuals, making anonymization steps unnecessary.

☑ 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?
Section 3.3.1 (Dataset Statistics and Distributions), Table 5, Table 6, and Appendix C.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix B (Finetuning 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?
Section 4.1, Section 4.3, Table 7, and Table 8. We report summary classification metrics (e.g., Accuracy, Precision, Recall, F1-score) and average confidence scores for a single evaluation run.

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.?
The task is relatively simple and clear, and does not require complicated explanations.

☒ 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 individuals listed are my friends and collaborators, and they all received appropriate compensation, but I did not manage the funds.

☒ 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 dataset is constructed exclusively from publicly available broadcast media, such as TV shows, movies, and public platforms like YouTube, TikTok, and Bilibili. Since the data features professional actors and public figures in broadcasted content, obtaining direct consent from the individuals depicted was not applicable and thus not discussed.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The paper relies on the annotation of publicly available media and secondary datasets. Formal Ethics Review Board (IRB) approval or exemption was not explicitly discussed or reported in the manuscript.

☑ **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 3.2.2. We disclosed the use of Large Language Models (LLMs) as an auxiliary tool to draft initial descriptions for emotion causes. Additionally, AI assistants were used strictly for copyediting and syntax optimization during the writing process.