Responsible NLP Checklist

Paper title: *A Culturally-diverse Multilingual Multimodal Video Benchmark & Model*
Authors: *Bhuiyan Sanjid Shafique, Ashmal Vayani, Muhammad Maaz, Hanoona Abdul Rasheed, Dinura Dissanayake, Mohammed Irfan Kurpath, Yahya Hmaiti, Go Inoue, Jean Lahoud, Md. Safirur Rashid, Shadid Intisar Quasem, Maheen Fatima, Franco Vidal, Mykola Maslych, Ketan Pravin More, Sanoojan Baliah, Hasindri Watawana, Yuhao Li, Fabian Farestam, Leon Schaller, Roman Tymtsiv, Simon Weber, Hisham Cholakkal, Ivan Laptev, Shin'ichi Satoh, Michael Felsberg, Mubarak Shah, Salman Khan, Fahad Shahbaz Khan*

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?
We discussed the Ethical Considerations and Risks in Section 8

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

☑ B1. Did you cite the creators of artifacts you used?
Yes, we have cited the creators of the artifacts throughout our paper including Section 3 and Section 4.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We discussed the License of the Artifacts including data sources, LMMs, and evaluation framework in Section N.

☑ 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)?
We discussed the Intended use of these Artifacts including data sources, LMMs, and evaluation framework in Section N.

☒ 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?
Our data does not explicitly contain personally identifiable information.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?

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.

We discussed the Documentation of the Artifacts including domains and subdomains for videos and questions in Section O.

☑ 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 have explained the dataset stats (Benchmark stats in Section B supplementary material and for training set, we have discussed it in Section 4).

☑ **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?
We discussed the Model Size, Computational Budget for training, evaluation, and ablation in Section P.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We discussed the Experimental Setup and Hyperparameters for training, evaluation, and ablation in Section Q.

N/A 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?
Although we have discussed our stats in detail throughout the paper and specially in Section 5, we did not plot error bars as it was out of domain for us. However, we did show a lot of qualitative samples in Section 5.

☒ 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?
We did not use any existing evaluation scores like ROUGE, we have used LLM as a judge so the existing was not required.

☑ **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.?
We have given detailed instructions to our human verifiers. We have summarized that in Section R.

☒ 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)?
None of our volunteers were paid in this project

☑ 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)?
We manually curated our cultural dataset with the help of human speakers.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our dataset is composed of existing subsets for the training and general part, for evaluation we constructed dataset from web-sources so ERB was not required.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
The demographics of the human annotators were discussed in detail in Section D in Supplementary Material.

☒ **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?
We did not use AI Assistants in our Research for coding or writing. We only did for evaluation which we have already mentioned above.