Responsible NLP Checklist

Paper title: *MedEBench: Diagnosing Reliability in Text-Guided Medical Image Editing*
Authors: *Minghao LIU, Zhitao He, Zhiyuan Fan, Qingyun Wang, Yi R. Fung*

> 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?
Yes, Section 11, Potential Risks

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

☑ B1. Did you cite the creators of artifacts you used?
Yes, section 2 and 3.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Yes, section 3.

☑ 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)?
Yes, section 10 Ethical considerations, Appendix B Details of MedEBench Dataset.

☑ 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?
Yes, section 2.1.1 Image Pair Collection, section 10 Ethical considerations.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Yes, section 10 Ethical considerations.

☑ 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?
Yes, section 2 MedEBench.

☑ **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?
Yes, section 3.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes, Appendix C Baseline Models Implementation Detail

☑ 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?
Yes, section 3.

☒ 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?
No, we did not use exsiting packages.

☑ **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.?
Yes, Appendix D, Human Evaluation and Validation Criteria.

☑ 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)?
Yes, Appendix D, Human Evaluation and Validation Criteria.

☑ 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)?
Yes, Appendix B Details of MedEBench Dataset.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No. The dataset is collected entirely from publicly available, open-access sources (e.g., Creative Commons images, published literature).

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No. We only reported the annotators qualifications (bachelors degrees) and recruitment source (crowdsourcing platform). To safeguard annotator privacy and due to the small number of annotators (two in total), we did not collect or disclose additional demographic or geographic details.

☑ **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 used AI assistants to support grammar checking and formatting.