Responsible NLP Checklist

Paper title: *Measure Twice, Click Once: Co-evolving Proposer and Visual Critic via Reinforcement Learning for GUI Grounding*
Authors: *Wenkai Wang, Xiyun Li, Hongcan Guo, Wenhao Yu, Tianqing Fang, Haitao Mi, Dong Yu, Shengyu Zhang*

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 paper includes a "Limitations" section (Section 6) discussing technical constraints such as inference latency and potential marker occlusion. However, it does not explicitly discuss broader societal or ethical risks.

☑ **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?
: Appendix A.1 explicitly states that the authors used pre-processed versions of the datasets where the original authors had already removed PII and offensive content.

☑ 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?
Appendix A.1 provides detailed statistics for the datasets used, such as reporting 21,750 unique UI screens for the Widget Captioning dataset and 9,800 pairs of screenshots and instructions for OmniACT

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The experimental setup is described in Section 5.1. Hardware details are provided in Appendix A.2 , and comprehensive training hyperparameters (e.g., learning rate, epochs, batch size) are detailed in Appendix A.5 and Table 4.

☒ 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?

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.

The main experimental results reported in Tables 1, 2, and 3 provide single accuracy scores (e.g., Top-1, Oracle@5) without error bars, variance, or statistical significance testing across multiple runs

☒ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

N/A 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.?
(left blank)

N/A 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)?
(left blank)

N/A 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)?
(left blank)

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
(left blank)

☑ **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?
Appendix F ("Information About Use Of Ai Assistants") states that large language models (ChatGPT, Copilot) were used to assist with coding and linguistic polishing, but no AI tools were used to generate scientific ideas.