Responsible NLP Checklist

Paper title: *Whats Missing in Vision-Language Models? Probing Their Struggles with Causal Order Reasoning*
Authors: *Zhaotian Weng, Haoxuan Li, Xin Eric Wang, Kuan-Hao Huang, Jieyu Zhao*

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.

N/A A2. Did you discuss any potential risks of your work?
This study only evaluates causal reasoning on publicly available image data and entails no additional risks.

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

☑ B1. Did you cite the creators of artifacts you used?
I used the publicly available, open-source datasets and I cited them in section 2, section 3, section 4. I also used some open-source models and I cited them in section2, section 3, section 4.

N/A B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
We used the publicly available, open-source datasets like VQA, VCR, LAION-400M, MSCOCO, which are widely used datasets in NLP research.

☑ 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)?
Section 2, section 3, section 4. We used the publicly available, open-source datasets like VQA, VCR, LAION-400M, MSCOCO, which are widely used datasets in NLP research.

N/A 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?
We only used the publicly available, open-source datasets like VQA, VCR, LAION-400M, MSCOCO, which are widely used datasets in NLP research that contain no personally identifying or offensive content.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Our benchmarks are constructed from the domain of VQA dataset and VCR dataset, and we have clarified this in section 2, section 3, section 4.

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.

☑ 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 reported the statistics for data in section 2, section 3, 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 reported the models and data size we used in section 2, section 3, section 4.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We have reported the experiments details in section 2, section 3, section 4.

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?
We conducted a single evaluation run per model.

N/A 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?
(left blank)

☑ **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.?
section 2.1. We used human annotators to annotate whether the text captions in our benchmarks are semantically coherent and fluent.

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)?
no external participants were recruited or compensated.

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)?
Not applicable. We only used the publicly available, open-source datasets like VQA, VCR, LAION-400M, MSCOCO, which are widely used datasets in NLP research.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. We only used the publicly available, open-source datasets like VQA, VCR, LAION-400M, MSCOCO, which are widely used datasets in NLP research.

N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
no external annotators were involved.

☒ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

N/A E1. If you used AI assistants, did you include information about their use?
We did not use AI assistants in our research.