Responsible NLP Checklist

Paper title: *DiscoSG: Towards Discourse-Level Text Scene Graph Parsing through Iterative Graph Refinement*

Authors: *Shaoqing Lin, Chong Teng, Fei Li, Donghong Ji, Lizhen Qu, Zhuang Li*

> 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?
This work improves evaluation methodology rather than creating deployable technology or safety-critical applications. The risk pathway from better evaluation metrics to societal harm is highly indirect and speculative.

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

☑ B1. Did you cite the creators of artifacts you used?
Section 2.1, 2.2

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Appendix B.5

☑ 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)?
Appendix B.5

☒ 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 consists primarily of GPT-4 generated captions describing visual scenes, which are less likely to contain PII or offensive content compared to human-authored text. The annotation task focused on extracting object relationships and visual attributes rather than personal information.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 2.4, Appendix C

☑ 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 2.4

☑ **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?
Appendix A.4

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix A.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?
Section 4

☑ 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?
Section 4

☑ **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.?
Appendix B.2

☑ 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)?
Appendix B.4

☑ 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)?
Appendix B.4

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No. The annotation task involved minimal risk as it focused on extracting object relationships from visual descriptions. Data collection was conducted by internal research team members rather than external human subjects, and involved analysis of existing publicly available research datasets.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No. The annotation was conducted by internal research team members (a senior researcher and computer science student) rather than a diverse annotator population. Detailed demographic characteristics were not collected as the annotators were co-authors participating in their research capacity.

☑ **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?
Yes. We used AI assistants for language editing and rephrasing to improve clarity and fluency of the manuscript text. AI was not used for generating research ideas, methodology, experimental design,

data analysis, or core content creation. All technical contributions, results, and conclusions are entirely the work of the authors.