Responsible NLP Checklist

Paper title: *VisualWebInstruct: Scaling up Multimodal Instruction Data through Web Search*
Authors: *Yiming Jia, Jiachen Li, Xiang Yue, Bo Li, Ping Nie, Kai Zou, Wenhu Chen*

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?
Our work involves web crawling and data collection from educational websites, which raises concerns about intellectual property rights and data usage permissions. We implemented compliance measures by avoiding websites with anti-crawling mechanisms and non-permissive licenses, but potential risks remain regarding copyright and fair use of educational content.

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

☑ B1. Did you cite the creators of artifacts you used?
We properly cited all datasets, models, and tools used in our research, including baseline models (GPT-4o, Gemini, MAmmoTH-VL), evaluation benchmarks (MMMU, MathVista, etc.), and technical frameworks. All references are included in our bibliography following standard academic citation practices.

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
While we mentioned compliance with copyright regulations and avoiding non-permissive licenses during data collection, we did not explicitly discuss the specific licenses for all artifacts used in our research. Our focus was primarily on technical implementation rather than detailed license analysis for each component.

☑ 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 used existing models (GPT-4o, Gemini, MAmmoTH-VL) and evaluation benchmarks for their intended research purposes. Our dataset will be publicly released to support multimodal reasoning research, consistent with academic data sharing practices.

☒ 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?

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.

Our dataset focuses on educational content from academic websites, homework platforms, and educational forums. We implemented filtering mechanisms to ensure educational value and removed irrelevant content, minimizing risks of personal information or offensive material.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4 provides comprehensive dataset statistics including domain distribution (Math 62.5%, Physics 14.5%, etc.), educational difficulty levels, and demographic coverage across multiple academic disciplines and grade levels.

☑ 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?
Table 3 details our pipeline statistics with 906K final QA pairs, 347K image-associated pairs, and 163K unique images. Section 4 provides complete breakdowns of data distribution, processing stages, and human evaluation results on 200 samples.

☑ **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?
Section 5.1 and Appendix C specify model parameters: MAmmoTH-VL2 is 7B parameters, using Qwen2.5-7B-Instruct as language tower and SigLip vision encoder. Appendix B reports pipeline costs totaling $10,771. Evaluation used 8 NVIDIA A100 80GB GPUs as detailed in Appendix D.3.

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
While we provided training configuration in Section 5.1 and detailed setup in Appendix C (learning rates, batch size, epochs), we did not conduct systematic hyperparameter search or report exploration of different hyperparameter values. We used standard configurations from prior work.

☒ 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?
Our results in Table 4 report single-run accuracy scores across benchmarks without error bars, confidence intervals, or multiple runs. We focused on demonstrating effectiveness rather than statistical robustness, following standard practice in multimodal model evaluation where single-run results are commonly reported.

☒ 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?
While we mentioned using evaluation frameworks like LMMsEval and specified some technical components (DeepSpeed Zero-3, mixed precision training), we did not provide comprehensive parameter settings for all preprocessing, normalization, or evaluation packages used in our pipeline.

☑ **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 4 mentions human evaluation on 200 randomly sampled QA pairs to assess dataset quality, but we did not provide the specific instructions, guidelines, or evaluation criteria given to human evaluators in the paper.

☒ 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)?
Our paper does not provide details about how human evaluators were recruited, their demographic information, compensation methods, or whether payment was adequate. The human evaluation component was mentioned only briefly without methodological details about participant management.

☒ 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)?
Our research involved web crawling from publicly available educational websites rather than direct data collection from individuals. We ensured compliance with website policies and avoided sites with anti-crawling mechanisms, but did not obtain explicit consent from original content creators since we used publicly accessible educational materials.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our study focused on automated web data collection from publicly available educational resources and did not involve direct human subjects research requiring IRB approval. The methodology centered on technical data processing rather than human participant studies.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
While we conducted human evaluation on 200 samples for quality assessment, we did not report demographic or geographic characteristics of the evaluators. The evaluation was limited in scope and focused on technical validation rather than comprehensive human subject analysis.

☑ **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?
Our research extensively used AI models (GPT-4o, Gemini-1.5-Flash) for data processing, QA extraction, consistency checking, and answer refinement throughout our pipeline. However, we did not explicitly discuss the role of AI assistants in the research process itself or manuscript preparation, focusing instead on technical methodology.