Responsible NLP Checklist

Paper title: *AIM-CoT: Active Information-driven Multimodal Chain-of-Thought for Vision-Language Reasoning*
Authors: *Xiping Li, Jianghong Ma*

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?
(left blank)

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

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?
(left blank)

☑ 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?
Relevant statistics for the datasets are reported in Section 5.1 and detailed in Appendix A.1. Specifically, Appendix A.1 provides descriptions and sizes for the M3CoT (11,459 samples) and ScienceQA (over 100,000 triples) benchmarks. Additionally, sample sizes for specific subsets used in ablation studies and qualitative analyses are reported in their respective sections, such as Sections (n=500) and Table 16 in Appendix I.4 (n=2318 for M3CoT subset).

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Yes, the experimental setup is outlined in Section 5.1, and the specific best-found hyperparameter values (such as candidate set size N_C and selection count K) are listed in Appendix A.4 (Table 7). Furthermore, we discuss the hyperparameter search and sensitivity analysis for the triggering threshold δ in Appendix G.1 and for the AVP module parameters in Appendix G.2.

☑ 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.

We report statistical significance tests (including p-values and test statistics) for our main performance comparisons in Section 5.2 (Table 1) and detail them in Appendix E (Table 10). Regarding descriptive statistics from sets of experiments, in Appendix I.2, we explicitly state that the source distribution experiments were repeated three times to ensure reliability, and we report the results for each individual run in Table 14 to demonstrate stability. Furthermore, Appendix M.2 (Table 21) reports the average inference time per instance to assess deployability.

☒ **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?
*We utilize AI assistants in this work for two primary purposes: **Writing Assistance:** We use AI tools across all sections to help identify grammatical errors and verify the correctness of LaTeX cross-references (e.g., '**??**' tags). **Experimental Evaluation:** We employ GPT-4v as a proxy judge in our qualitative and ablation analyses. Specifically, in Section 5.6, GPT-4v evaluates the alignment of selected visual regions with human intuition by assessing their relevance, completeness, and helpfulness. In Appendix L.1, GPT-4v serves as an external evaluator to detect factual inconsistencies and calculate the hallucination rate of the generated image descriptions.*