Responsible NLP Checklist

Paper title: *Textual Steering Vectors Can Improve Visual Understanding in Multimodal Large Language Models*
Authors: *Woody Haosheng Gan, Deqing Fu, Julian Asilis, Ollie Liu, Vatsal Sharan, Robin Jia, Willie Neiswanger*

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?
Section "Ethical Considerations" (appears after Conclusion and Limitations section, before Reference and Acknowledgement).

☑ **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?
We use publicly available benchmark datasets (CV-Bench, What'sUp, BLINK, CLEVR, Super-CLEVR, VQAv2, COCO Captions, DocVQA, ChartQA, VTabFact) that have been widely used in prior research and do not contain personally identifying information. For steering vector extraction (Section 4.1), we created small sentence-anchor pairs focusing on visual concepts (spatial relationships, counting, attributes, entities) that contain no personal or 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?
Section 5.1 reports dataset statistics for CV-Bench (2,638 datapoints with train/test splits, i.e, 150 for test split on each task). Section 6.1 reports statistics for out-of-distribution datasets including What'sUp-A, What'sUp-B, BLINK, CLEVR, and Super-CLEVR, along with validation subset sizes. Appendix C.2 provides statistics for additional evaluation datasets (VQAv2, COCO Captions, DocVQA, ChartQA, VTabFact). Section 4.1 and Table 3 in Appendix A.1 describe the sentence-anchor pairs used for steering vector extraction.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5.1 describes the experimental setup including model selection, grid search procedure

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.

for hyperparameters (layer and scale factor), and the specific hyperparameter ranges tested: = {0.1,0.2,0.4,0.6,0.8,1.0} for MeanShift and {10,20,30,40,50,60} for SAE/Probe. Layer ranges are also specified for each model in Section 5.1 and Appendix C.2. Figure 4b and 5 show best-found hyperparameters.

☑ 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?
Tables 1, 2, 4, 5, 6, and 7 report accuracy scores with statistical significance testing via bootstrap resampling (10,000 iterations, $p < 0.05$) marked with stars. Tables in Section 6 and Appendix report mean improvements across multiple datasets and models. All results represent single deterministic runs (greedy decoding) with the best hyperparameters found via grid search.

☒ **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 used o3-mini to verify the alignment of SAE features with visual concepts during steering vector extraction, as described in Appendix A.1 (Algorithm 1 and verification prompt template). We also used GPT-4o to generate candidate prompts for the prompting baseline method (Section 4.2 and Appendix A.2). All AI-assisted components are explicitly documented in the methodology sections