Responsible NLP Checklist

Paper title: *MagicBench: Diagnosing Visual Agency Loss and Semantic Dependency in Multimodal LLMs*
Authors: *Tang Da Huang, Weidong Tang, Wen Qi Xu, Xianpeng Guo*

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?
Potential robustness concerns and the broader implications of cross-modal interference for downstream applications are addressed in Section 6.3 and the Limitations section.

☑ **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 manually screened all 402 videos in MagicBench to ensure they contain no personally identifying information or offensive content. Since these are professional magic tutorials intended for public distribution, the only identifiable information is the public identity of the performers. This is addressed in Section 3.1.

☑ 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?
Detailed statistics for MagicBench, including the number of samples per linguistic interference type (N=402 total), video sampling rates, and Physical Constraint Set complexity metrics, are reported in Section 3.1 and Section 3.2. Comprehensive per-model performance breakdowns are provided in Table 7.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The experimental setup, including the specific MLLM versions, inference temperature (T=0.6), and prompt configurations for both Vision-Only and Multimodal settings, is detailed in Section 4.1 (Experimental Setup). Full prompt templates are provided in Table 8 in the Appendix.

☑ 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 mean scores along with standard errors estimated via 1,000 bootstrap iterations to ensure statistical reliability. Significance levels (e.g., $p < 0.01$) are explicitly marked using the symbol in Table 3 and Table 5, as discussed in Section 4.2.

☑ **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.?
The annotation protocol and specific criteria for constructing the Physical Constraint Set (PCS) are detailed in Section 3.1 and further formalized using First-Order Logic in Appendix D. The scoring rubric for human validation is provided in Table 1.

☒ 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)?
All data annotation and the human validation audit (N=80) were conducted internally by the authors as part of their academic research duties. No external crowdworkers or experimental participants were recruited or paid.

☒ 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)?
The dataset comprises publicly available professional magic tutorial videos intended for public viewing. Since the annotations were performed internally by the authors, formal participant consent forms were not required.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our study involves the expert analysis of publicly available instructional videos and does not involve human intervention, the collection of private personal information, or behavioral experiments on human subjects. Therefore, it was determined to be exempt from ethics review board approval.

☑ **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?
AI assistants were used exclusively for language polishing, grammatical refinement, and style consistency of the manuscript. All core scientific ideas, experimental designs, and results were generated solely by the human authors.