

Responsible NLP Checklist

Paper title: *From Mimesis to Metamorphosis: Evolving VLM Judges via In-Context Comparing and Knowledge Internalization*

Authors: *Juntuo Wang, Yuming Qiao, Yifan Yang, Lunxi Yuan, Liang Luo, Dan Meng*

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 focuses on the methodological framework for improving the scale-consistency and alignment of vision-language models in subjective evaluation tasks. It primarily utilizes publicly available, standard research datasets (e.g., PARA, AVA) and does not involve human subject experiments or the collection of sensitive personal data. While all subjective assessment models are subject to the inherent biases of their training data, this study does not introduce novel ethical risks or applications that pose direct social harm beyond the existing norms in the field of image quality and aesthetics research.

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?

Appendix B.

C. Did you run computational experiments?

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

Section 4.1 (Implementation Details) and Appendix B. Hyperparameters were determined empirically based on established training protocols for 7B-scale vision-language models and hardware constraints (8 A100). The specific values reported represent the best configuration used for all reported results.

- 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,

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.

etc. or just a single run?

We report descriptive statistics including the mean and standard deviation for our model's predicted scores in Appendix C.

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.?

(left blank)

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)

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)

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?

I only use AI assistants to help check grammar issues.