Responsible NLP Checklist

Paper title: *Selective Test-Time Debiasing for CLIP via Reward Gating*
Authors: *Jaeho Han, Jisoo Yang, Hyeondong Woo, Mingyu Jeon, Sunjae Yoon, Junyeong Kim*

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?
See Limitations and Ethics Statement sections. We discuss risks including potential bias propagation from the external reward model (CLIP ViT-L/14), the implicit assumption of a uniform target distribution that may not suit all domains, sensitivity to hyperparameters and attribute estimators, and the careful handling required for sensitive demographic data. We recommend auditing reward sources and avoiding high-stakes decision-making without additional validation.

☑ **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?
See Ethics Statement. We use publicly available fairness benchmarks (FairFace, UTKFace, FACET) that contain face images and protected-attribute annotations (gender, age, race). Our work does not aim to identify individuals, and we use these datasets strictly in accordance with their original licenses and intended usage conditions for fairness evaluation. We rely on the anonymization and curation procedures provided by the original dataset releases, and we do not collect, release, or redistribute any new personally identifying information.

☑ 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 4.1 (Datasets) and Appendix A.5. We use the standard public splits of FairFace (validation set), UTKFace, FACET, ImageNet-1K, and Flickr1k (1,000 image-text pairs) without modification, following their original release protocols. The reference attribute set used for gating is constructed with M=5 images per attribute class sampled from UTKFace, as described in Appendix A.5. Since we do not create new datasets or alter existing splits, we refer readers to the original dataset papers for full statistics.

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

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

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.

See Section 4.4 (Implementation Details) and Appendix A. We report the backbones used (CLIP ViT-B/16, ViT-B/32, ViT-L/14), optimizer (AdamW with learning rate 1e-4, weight decay 0), reward strength =1000, Gaussian kernel temperature =0.25, gating threshold =0.02, and adaptive computational budget (T=3 steps with K=10 candidates when (q)=0; T=10 steps with K=1024 candidates when (q)=1). Sensitivity analyses for the candidate budget K, debiasing strength , and gating threshold are provided in Section 5.1 (Figures 4 and 5). Bias subspace construction details and prompt templates are described in Appendix A.5.

☒ 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?
We report results from a single run for each configuration without error bars. RG-TTA performs episodic test-time adaptation with parameter resets after each query, and evaluation uses fixed public benchmark splits, which makes the procedure largely deterministic across runs. Per-query runtime statistics in Table 5 are reported as averages over 1,000 queries. We acknowledge that adding multi-seed variance estimates would further strengthen the reporting and leave this for an extended version.

☒ **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.?
N/A. This work does not involve human annotators or human subjects; all experiments use existing publicly available benchmark datasets.

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)?
N/A. This work does not involve human participants or annotators; no recruitment or payment was required.

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)?
N/A. We did not collect new data from human subjects. All experiments use publicly available benchmark datasets (FairFace, UTKFace, FACET, ImageNet-1K, Flickr1k) under their original licenses and intended usage conditions, as discussed in the Ethics Statement.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
N/A. This work did not involve new data collection from human subjects, so ethics review board approval was not applicable. All experiments use publicly available benchmark datasets under their original licenses.

☑ **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 (e.g., ChatGPT/Claude) were used solely to assist with language polishing, grammar checking, and improving the clarity of writing. They were not used for generating research ideas, designing experiments, producing experimental results, or writing code. All technical content, methodology, and analyses are the authors' own work, and the authors take full responsibility for the final content.