Responsible NLP Checklist

Paper title: *Omni-Embed-Audio: Leveraging Multimodal LLMs for Robust Audio-Text Retrieval*
Authors: *HaeJun Yoo, Yongseop Shin, Insung Lee, Myoung-Wan Koo, Du-Seong Chang*

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 the Limitations section. We discuss several potential risks and failure modes of the work, including dependence on native-audio multimodal LLM backbones, higher deployment memory costs than compact CLAP models, residual bias in hard-negative mining, and possible synthetic bias or under-coverage in LLM-generated UIQs.

☑ **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?
We used public audio-text datasets centered on environmental or event-centric audio retrieval (e.g., AudioCaps, Clotho, MECAT, and WavCaps subsets) rather than datasets intended to identify individuals. The work does not involve collecting or releasing personal profiles, and the paper does not focus on offensive-content modeling. Accordingly, this item is not central to the present study.

☑ 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?
Yes. Relevant dataset statistics and split information are reported in Section 4.1 (Datasets), and UIQ benchmark statistics are reported in Section 3.2.3. Additional provenance and overlap statistics are provided in Appendix B (Data Leakage Analysis).

☑ **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 and training configuration are reported in Section 4.2 (Models and Training), and implementation details including LoRA configuration, embedding projection, and contrastive objective are provided in Appendix A (Implementation Details).

☑ 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?
Yes. We report retrieval metrics (R@1, R@5, R@10) across datasets in Tables 24, and explicitly indicate when results are reported as means across datasets. For UIQ validation, we also report descriptive statistics including means and standard deviations in Table 1 and Appendix C. Additional detailed per-dataset and mean results are reported in the appendices.

☑ **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.?
Yes. The human evaluation question, rating scale, and interface description are reported in Appendix E.1 (Human Evaluation). The benchmark-generation prompts are separately documented in Appendix D, while the participant-facing evaluation prompt and rating procedure are described in Appendix E.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)?
Yes. Recruitment and compensation details are reported in Section 3.2.3. Annotators were recruited through an internal Sogang University community posting open to undergraduate and graduate students, and were paid an hourly wage.

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)?
This work uses existing public audio-text datasets for retrieval benchmarking and does not curate personal data about identifiable individuals. The human study concerns query-validity ratings over public examples rather than collection of personal participant data for release; thus this item is not central to the present work.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No. The human evaluation was a minimal-risk validation task in which annotators rated the naturalness of search queries for public audio-caption examples. The study did not involve sensitive personal data, medical interventions, or vulnerable populations.

☑ **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?
Yes. AI systems are explicitly documented as part of the research pipeline. We describe the use of GPT-5.1 for UIQ generation in Section 3.2.2 and Appendix D, and the use of Claude Opus 4.5 for LLM-based evaluation in Section 3.2.3 and Appendix E.2. All generated outputs were checked by the authors and, where applicable, complemented by human validation.