Responsible NLP Checklist

Paper title: *LaMI: Augmenting Large Language Models via Late Multi-Image Fusion*
Authors: *Guy Yariv, Idan Schwartz, Yossi Adi, Sagie Benaim*

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 8 "Ethical Considerations" (page 6) discusses hallucinations and factual grounding, bias propagation from component models (LLM, CLIP, SDXL-turbo).

☑ 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?
Sections A.1, A.2, and A.3 (Appendix, page 9) report dataset details: Visual Genome Regions (5.4M images), Laion-220K, 2% of Wikitext-103 for training, and full benchmark listings with evaluation protocols for all test sets. Section A.2 details the 3,200 yes/no questions used for debiasing.

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section A.1 (page 9) reports: CLIP-ViT-B/32 as vision encoder, SDXL-turbo for generation, two-stage training pipeline on 4x A100 GPUs, 40K iterations (batch 256, lr 5e-4) + 10K fine-tuning iterations (batch 128, lr 5e-5), AdamW optimizer, GPU hours per model (192 for Llama-3, 90 for Gemma-2B, 50 for GPT-2), k=6 images at inference.

☒ 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?
(left blank)

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.

☒ **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?
ChatGPT/Claude was used for grammar fixing assistance during writing.