Responsible NLP Checklist

Paper title: *Language-Grounded Multi-Domain Image Translation via Semantic Difference Guidance*
Authors: *jongwon ryu, Joonhyung Park, Jaeho Han, Yeong-Seok Kim, Hye-Rin Kim, 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?
functionally same as stable diffusion model

☑ **B. Did you use or create scientific artifacts? (e.g. code, datasets, models)**

☑ B1. Did you cite the creators of artifacts you used?
ection 4 (Experiments) and the References cite all datasets and models used, including CelebA, CelebA-Dialog, BDD100K, Animal Faces, CLIP, DINOv2, and Stable Diffusion v2.1.

☒ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
All datasets used (CelebA-Dialog, BDD100K) are publicly available for non-commercial academic research.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 4 (Experiments) all datasets are used within their intended research purposes for image-to-image translation and attribute editing.

☑ 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?
Section 4 (Experiments) only publicly available research datasets were used. These datasets contain no personally identifiable information (PII) or offensive content.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 4 (Experiments) and the Appendix describe dataset composition, attribute combinations, and preprocessing procedures.

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.

☑ 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 (Experiments) includes dataset sizes, attribute counts, and splits for 1-, 2-, and 3-domain settings.

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

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 4.1 (Implementation Details) the model is based on Stable Diffusion v2.1. Training was conducted on two NVIDIA A6000 GPUs for 200k steps.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1 (Implementation Details) reports the learning rate, scheduler, guidance scale, and DDIM inversion settings.

☑ 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?
Section 4.2 (Quantitative Evaluation) reports FID, CLIP-FID, SSIM, and LPIPS metrics. Human evaluation results (mean scores) are added in the revision.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation, such as NLTK, SpaCy, ROUGE, etc.), did you report the implementation, model, and parameter settings used?
Section 4 (Experiments) the implementation uses PyTorch and HuggingFace Diffusers.

☑ **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.?
Section 4 (Experiments) Evaluators were instructed to rate attribute correctness and naturalness on a scale of 0-10.

☑ 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)?
Five human evaluators were recruited to assess 100 samples across four datasets.

☑ 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)?
Participants were informed that their ratings would be used for academic evaluation of image translation quality.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The human evaluation involved non-sensitive rating of generated images and did not require formal IRB approval in this context.

☒ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Demographic details of the five evaluators were not recorded as the task focused on objective image quality and semantic alignment.

☑ **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 was used solely for English language polishing and improving the readability of the manuscript. No technical content, experimental design, or data analysis was generated by AI.