Responsible NLP Checklist

Paper title: *X-CoT: Explainable Text-to-Video Retrieval via LLM-based Chain-of-Thought Reasoning*
Authors: *Prasanna Reddy Pulakurthi, Jiamian Wang, MAJID RABBANI, Sohail Dianat, Raghuveer Rao, Zhiqiang Tao*

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 "Limitations". We discuss that annotation quality depends on LLM reasoning, and retrieval may fail in noisy or domain-specific scenarios. Potential risks include bias in automatically generated annotations and possible dataset noise propagation.

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

☑ B1. Did you cite the creators of artifacts you used?
Section 4. We cite the original benchmark datasets (MSR-VTT, MSVD, LSMDC, DiDeMo) and prior baseline methods.

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
Section 3.2 and GitHub repository. We release code and annotations under an open research license (following the terms of the original datasets, which are for research use only).

☑ 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 3.2. The additional annotations are designed for research purposes only, consistent with the intended academic use of the original datasets.

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)

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3.2 and Appendix. We describe the annotation pipeline, filtering steps, and provide documentation with code and data on GitHub.

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 and Appendix. We report dataset statistics (number of annotations, noisy vs. complete tags, robustness tests).

☑ **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?
We report runtime, GPU memory, and infrastructure details in the Appendix.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4. Hyperparameters and experimental setup are described.

☑ 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. Results are reported with mean/median ranks, recall@K, ablation comparisons, and robustness studies.

☑ 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 3 and Appendix A. We describe use of ResNet18 (ImageNet pretrained), CLIP, X-Pool, and VLM2Vec for preprocessing, baseline comparisons, and evaluation.

☒ **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)

N/A D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
(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?
Section 3.2. We used an open-source MLLM (Qwen2.5-VL-7B-Captioner-Relaxed) to automatically generate frame-level captions and structured video annotations as part of the dataset curation pipeline. No AI assistants were used in writing the manuscript.