Responsible NLP Checklist

Paper title: *QEVA: A Reference-Free Evaluation Metric for Narrative Video Summarization with Multimodal Question Answering*
Authors: *Woojun Jung, 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?
Our work does not pose direct societal or ethical risks, as it introduces an evaluation metric (QEVA) and an annotated dataset for video summarization. The contributions are methodological and evaluative in nature, and do not involve deploying generative models for end-user applications. Potential risks such as model misuse, bias, or harmful outputs are minimal in this context.

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

☑ B1. Did you cite the creators of artifacts you used?
We cited the creators of all datasets (MLVU) and baseline metrics/models used in our experiments (see Section 2: Related Work). Our own artifacts (QEVA and MLVU-VS-Eval) are clearly described and referenced in Section 3 and Section 4

☑ B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
The datasets and models we used (e.g., MLVU benchmark, Video-LMMs such as Qwen2.5-VL, InternVL, Video-LLaVA, GPT-4o) were all employed in accordance with their intended research licenses, as discussed in Section 4.1. For our own released artifacts (QEVA code and MLVU-VS-Eval dataset), we provide open access under a research-permissive license (see supplementary material).

☑ 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)?
All artifacts were used solely for research purposes, consistent with their intended use and access conditions. We did not create derivative datasets outside the research context. This is discussed in Section 4.1 (MLVU-VS-Eval) and Section 5 (Conclusion, future research).

☒ 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?

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.

Our work does not involve personally identifiable information (PII) or offensive content. The dataset (MLVU benchmark and our derived MLVU-VS-Eval) consists of publicly available video content and human-written summaries, all anonymized and free of sensitive personal data.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
We documented the proposed dataset (MLVU-VS-Eval) in Section 4.1, including domain coverage, annotation procedures, and evaluation criteria. We also provided detailed prompts and methodology for QEVA in Appendix A.

☑ 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?
We reported dataset statistics in Section 4.1. Specifically, MLVU-VS-Eval consists of 200 videos (~15 minutes average length) and 800 generated summaries, each annotated by two human evaluators. We also provided inter-annotator agreement (Krippendorffs = 0.68) and details of evaluation splits.

☑ **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 reported model choices (e.g., Gemini-1.5 Pro, GPT-4o, Qwen2.5-VL, InternVL3, Video-LLaVA) and experimental infrastructure in Section 3.5 and Section 4.4. While we did not provide exact GPU hours, we described the computational setup and discussed cost considerations (API usage) as a limitation.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
We detailed prompts, hyperparameters, and implementation settings for QEVA in Section 3.5 and Appendix A, ensuring reproducibility of the evaluation pipeline.

☑ 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 reported descriptive statistics including dataset sizes (200 videos, 800 summaries), inter-annotator agreement (Krippendorffs = 0.68), and summary statistics of correlation results across metrics (Tables 14, Section 4).

☑ 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?
We specified implementation details of packages and metrics, including baseline metrics (BLEU, ROUGE, METEOR, CIDEr, SPICE, BERTScore, CLIPScore) and QEVAs QA-based setup (see Section 2.1, Section 3.5, and Section 4).

☑ **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.?
Annotators received clear instructions regarding the evaluation criteria (Coverage, Factuality, Temporal Coherence) and example guidelines (Section 3.1 and Appendix 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)?
Annotators were voluntarily recruited from our institution and fairly compensated for their time (see Ethics Statement)

☑ 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)?
Informed consent was obtained from all annotators before participation, and they were provided with a clear description of the studys objectives and tasks (Ethics Statement).

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Our institutions review board determined that formal IRB approval was not required, as the study involved subjective evaluations of anonymized system outputs (Ethics Statement).

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Annotators were graduate and undergraduate students in computer science and related fields from our institution (Section 4.1).

☑ **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?
Our use of AI assistants (e.g., ChatGPT) was limited to improving the clarity of writing, language editing, and organizing checklist responses. No experimental design, implementation, or results were generated by AI tools.