Responsible NLP Checklist

Paper title: *VideoPASTA: 7K Preference Pairs That Matter for Video-LLM Alignment*
Authors: *Yogesh Kulkarni, Pooyan Fazli*

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.

N/A A2. Did you discuss any potential risks of your work?
(left blank)

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

☑ B1. Did you cite the creators of artifacts you used?
We cite creators of datasets used (e.g., ActivityNet in Section 4), foundational models (e.g., Qwen2.5-VL, InternVL2.5 in Section 4 and Table 1), and relevant software frameworks (e.g., SWIFT, LMMs-Eval in Section 4).

N/A B2. Did you discuss the license or terms for use and/or distribution of any artifacts?
(left blank)

☑ 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)?
Existing artifacts (datasets, models) were used for their intended research purposes in video understanding and model alignment. Our created preference data is for the intended use of improving Video-LLM alignment as described.

☒ 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 utilize publicly available research datasets (e.g., ActivityNet) which are standard in the field. Our method generates preference pairs based on model responses to video content and does not involve collecting new data with PII or direct human annotation.

N/A B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
(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.

☑ 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?
Key statistics for our generated preference dataset (initial count, filtering process, final count of 7,020 pairs) are in Section 4 and Appendix E. Details of benchmark datasets used for evaluation are standard and cited in Section 4.

☑ **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?
Sizes of foundational models (e.g., 7B, 8B) are specified in Section 4 and Table 1. The training setup (four NVIDIA L40S GPUs 48GB each, 32-frame limit) is described in Section 4. LoRA fine-tuning on 7k pairs for one epoch typically takes approximately 6~12 hours depending on the target model.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
The experimental setup, including the DPO process, LoRA parameters, DPO scaling factor, and loss weights, is detailed in Section 4 and Section 3.5. Further ablations on hyperparameters are in Appendix D.

☑ 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?
Results in tables are from single runs on established benchmarks, following common practice. We report mean scores where provided by the benchmark (e.g., TempCompass Avg.). Standard deviations or error bars are not reported as our focus is on improvements via a deterministic fine-tuning process from specific checkpoints.

☑ 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 use LMMs-Eval for standardized evaluation and SWIFT for training, as cited in Section 4. Specific parameter settings for these packages are their standard configurations.

☒ **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?**

N/A E1. If you used AI assistants, did you include information about their use?
(left blank)