Supplementary Material for LongVLM: Efficient Long Video Understanding via Large Language Models

Yuetian Weng[1], Mingfei Han[3,4], Haoyu He[1], Xiaojun Chang[2,3], and Bohan Zhuang[1†]

[1]ZIP Lab, Monash University, Australia [2]School of Information Science and Technology, University of Science and Technology of China
[3]Department of Computer Vision, MBZUAI [4]ReLER, AAII, UTS
{yuetian.weng,haoyu.he}@monash.edu, hmf282@gmail.com, xjchang@ustc.edu.cn, bohan.zhuang@gmail.com

We organize our supplementary material as follows.

- In Sec. A, we provide additional ablation results.
- In Sec. B, we provide more generation results from the proposed LongVLM.

A Additional Results

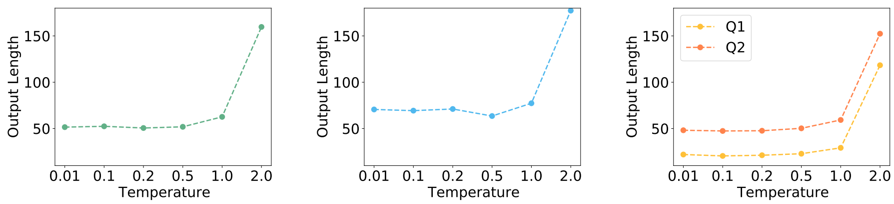
Effects of temperature. We show the effects of temperature settings in the LLM on both performance scores and averaged word counts in responses on Video-ChatGPT benchmark [2]. As demonstrated in Tab. A and Fig. A, lower temperatures lead to shorter but more accurate responses, with tokens that are more deterministic and closely aligned with the most relevant predictions. However, increasing the temperature from 0.2 to 2.0 lowers performance and raises word counts, leading to longer, more creative responses that may be irrelevant and inaccurate. Therefore, we set the temperature at 0.2 in our model.

Table A: Effects of temperature. We report the Correctness of Information (**CI**, Temporal Understanding (**TU**), and Consistency (**C**) by varying temperature in {0.01, 0.1, 0.2, 0.5, 1.0, 2.0}.

Temperature	0.01	0.1	0.2	0.5	1.0	2.0
CI	2.74	**2.76**	**2.76**	2.69	2.35	1.21
TU	2.22	2.16	**2.39**	2.20	1.99	1.25
C	2.85	2.89	**3.11**	2.62	2.21	1.01

More results on zero-shot QA. We evaluate our model in a zero-shot manner on subsets of two datasets: (1) Egoschema [3], a long-form video QA dataset derived from Ego4D [1]; and (2) MAD [4], an audio description (AD) generation benchmark that requires understanding long contexts in hour-long movies. We sample videos at 1 fps for Egoschema and 5 fps for MAD. We report the accuracy

† Corresponding author.



(1) CI **(2) TU** **(3) C**

Fig. A: Average word counts in responses for various temperature settings. "Q1" and "Q2" denote two questions addressing similar perspectives in the Consistency (C) evaluation metrics.

Table B: Model performance on Egoschema and MAD.

Method	Egoschema	MAD
VideoChat	46.6	1.90
Video LLaMA	38.8	1.86
Video-ChatGPT	49.6	1.93
BT-Adapter	54.6	2.14
LongVLM	**57.6**	**2.21**

Table C: Ablation of the local and global aggregation design.

Variants	ANET-QA	Egoschema
Pooling	45.4	48.0
Merging	46.5	52.4
[L, G]	47.0	56.0
[G, L]	**47.6**	**57.6**

(%) for Egoschema and the response score (ranging from 0 to 5) for MAD in Tab. B, demonstrating the effectiveness of the proposed LongVLM on longer video datasets.

Additional ablation results on zero-shot QA. We verify the local feature aggregation and global semantics integration strategies on ANET-QA [6] and Egoschema [3]. We report the accuracy (%) in Tab. C. The results demonstrate the effectiveness of token merging module compared to the pooling operation on longer videos. Additionally, concatenating global features before local features yields better results than the opposite concatenation order.

B Additional Qualitative Examples

We provide more generated examples from VideoChatGPT benchmark, ANET-QA and MSRVTT-QA. The proposed LongVLM can generate precise and accurate responses that correspond to video content and correctly answer questions about detailed information in videos.

Fig. B: Additional response examples from the Video-ChatGPT benchmark [2] generated by our LongVLM, focusing on temporal understanding (**TU**) evaluation perspective. Text in bold indicates the correct content.

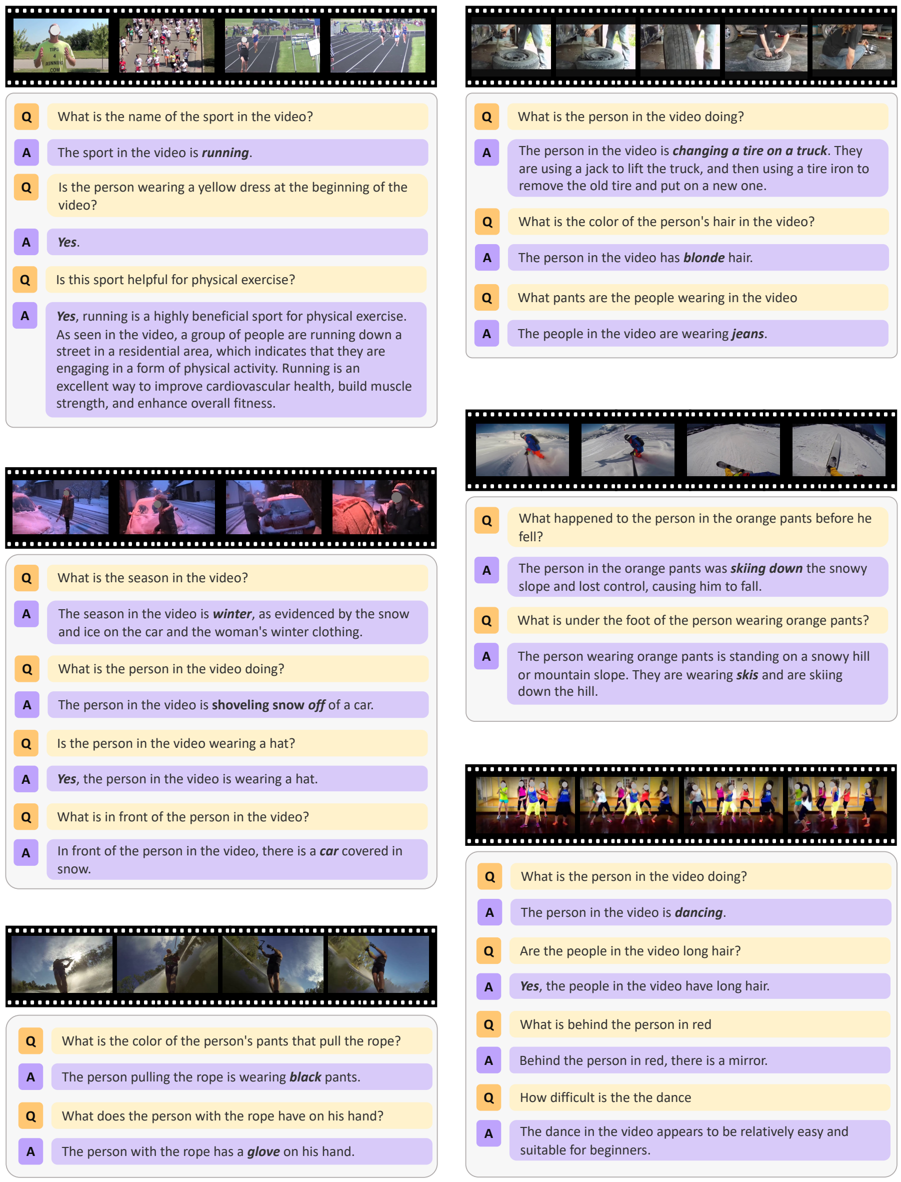


Fig. C: Additional examples from Zero-shot ANET-QA [6] generated by the proposed LongVLM. Text in bold indicates the correct content.

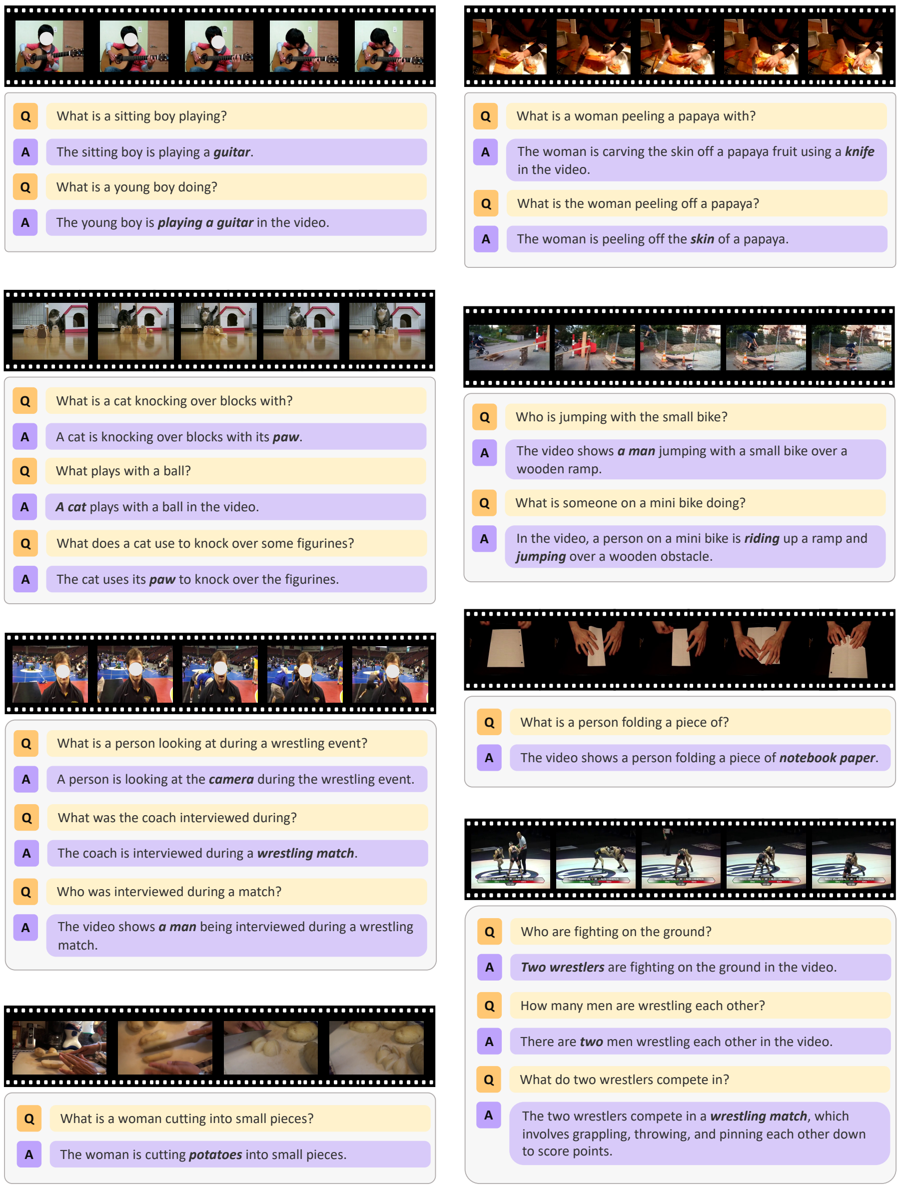


Fig. D: Response examples from Zero-shot MSVD-QA and MSRVTT-QA [5] generated by the proposed LongVLM. Text in bold indicates the correct content.

References

1. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR. pp. 18995–19012 (2022) 1
2. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424 (2023) 1, 3
3. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. NIPS **36** (2024) 1, 2
4. Soldan, M., Pardo, A., Alcázar, J.L., Caba, F., Zhao, C., Giancola, S., Ghanem, B.: Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In: CVPR. pp. 5026–5035 (2022) 1
5. Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: ACMMM. pp. 1645–1653 (2017) 5
6. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: AAAI. vol. 33, pp. 9127–9134 (2019) 2, 4