

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

Yuetian Weng¹, Mingfei Han^{3,4}, Haoyu He¹,
Xiaojun Chang^{2,3}, and Bohan Zhuang^{1†}

¹ZIP Lab, Monash University, Australia ²School of Information Science and
Technology, University of Science and Technology of China

³Department of Computer Vision, MBZUAI ⁴ReLER, AAIL, 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) Egoscema [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 Egoscema and 5 fps for MAD. We report the accuracy

[†] Corresponding author.

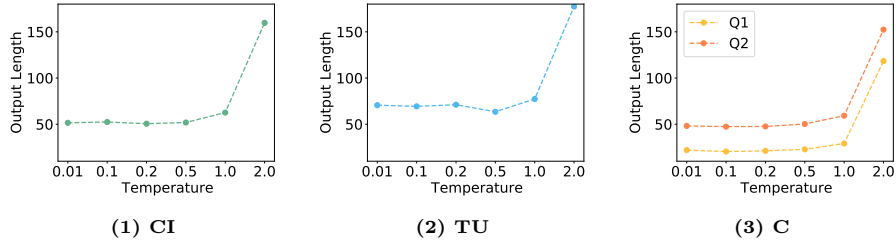


Fig. A: Average word counts in responses for various temperature settings. "Q1" and "Q2" denote two questions addressing similar perspectives in the Consistency (C) evaluation metrics.

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 MSRVT-QA. The proposed LongVLM can generate precise and accurate responses that correspond to video content and correctly answer questions about detailed information in videos.



Question What is the first maneuver the man performs on the parallel bars and what is the final maneuver he performs before attempting a backswing?

LongVLM The first maneuver the man performs on the parallel bars is a **handstand**. He then performs a backswing, which involves swinging his body backward and upward while holding onto the bars. The final maneuver he performs before attempting a backswing is a **handstand on the bars**.




Question What is the sequence of events in the cheer squad routine shown in the video?

LongVLM The **cheer squad** routine in the video consists of a series of movements and stunts, showcasing the team's synchronization and athleticism. The routine starts with the cheerleaders standing in a formation, and then they begin to perform various stunts, with some girls performing **jumps** and **flips** while others are holding hands. The routine culminates with the cheerleaders **performing a pyramid formation**, where they are stacked on top of each other. Throughout the routine, the cheerleaders maintain a high level of energy and enthusiasm, which is essential for a successful performance.



Question What happens after the tattoo artist marks the girl's ear?

LongVLM After the tattoo artist marks the girl's ear, they begin to **tattoo** her ear. The video shows the artist working on the girl's ear, and it appears that they are **using a needle to create the tattoo design**. The process involves the artist **making incisions** in the skin, **adding color**, and possibly **shading to create the desired tattoo**. The girl is **lying on her back**, and the artist is working on her ear, which is a common position for tattooing. The tattooing process **can be painful**, and the girl may need to endure some discomfort during the process. After the tattooing is complete, the artist will likely **clean the area and apply a bandage or dressing** to protect the tattoo and prevent infection.



Question How does the man iron the shirt and what does he do afterwards in the video?


LongVLM The man in the video is ironing a shirt using an ironing board. He **first places the shirt on the ironing board** and then **uses the iron to press it**. After ironing the shirt, he **takes it off the ironing board and hangs it up**. The entire process is captured in the video, providing a clear demonstration of how to iron a shirt using an ironing board.



Question What happens towards the end of the video?

LongVLM At the end of the video, the windsurfer **falls off the board and into the water**.

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.



Q What is the name of the sport in the video?


A The sport in the video is **running**.

Q Is the person wearing a yellow dress at the beginning of the video?

A Yes.

Q Is this sport helpful for physical exercise?

A Yes, running is a highly beneficial sport for physical exercise. As seen in the video, a group of people are running down a street in a residential area, which indicates that they are engaging in a form of physical activity. Running is an excellent way to improve cardiovascular health, build muscle strength, and enhance overall fitness.



Q What is the person in the video doing?

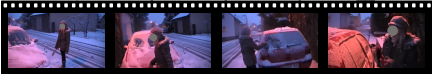
A The person in the video is **changing a tire on a truck**. They are using a jack to lift the truck, and then using a tire iron to remove the old tire and put on a new one.

Q What is the color of the person's hair in the video?

A The person in the video has **blonde** hair.

Q What pants are the people wearing in the video?

A The people in the video are wearing **jeans**.



Q What is the season in the video?

A The season in the video is **winter**, as evidenced by the snow and ice on the car and the woman's winter clothing.

Q What is the person in the video doing?


A The person in the video is **shoveling snow off** of a car.

Q Is the person in the video wearing a hat?

A Yes, the person in the video is wearing a hat.

Q What is in front of the person in the video?

A In front of the person in the video, there is a **car** covered in snow.




Q What happened to the person in the orange pants before he fell?

A The person in the orange pants was **skiing down** the snowy slope and lost control, causing him to fall.

Q What is under the foot of the person wearing orange pants?

A The person wearing orange pants is standing on a snowy hill or mountain slope. They are wearing **skis** and are skiing down the hill.



Q What is the person in the video doing?

A The person in the video is **dancing**.

Q Are the people in the video long hair?


A Yes, the people in the video have long hair.

Q What is behind the person in red?

A Behind the person in red, there is a mirror.

Q How difficult is the the dance

A The dance in the video appears to be relatively easy and suitable for beginners.



Q What is the color of the person's pants that pull the rope?

A The person pulling the rope is wearing **black** pants.

Q What does the person with the rope have on his hand?

A The person with the rope has a **glove** on his hand.

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

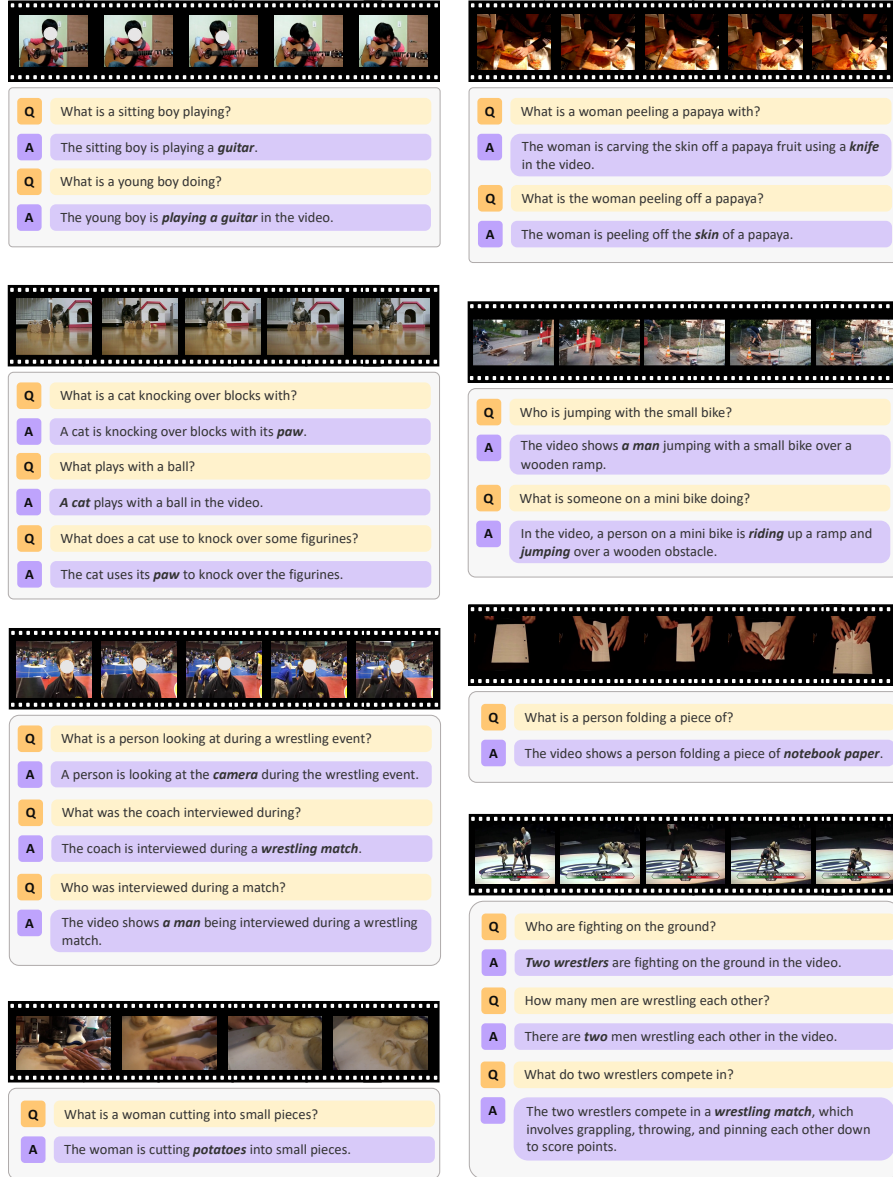


Fig. D: Response examples from Zero-shot MSVD-QA and MSRVT-QA [5] generated by the proposed LongVLM. Text in bold indicates the correct content.

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](#)