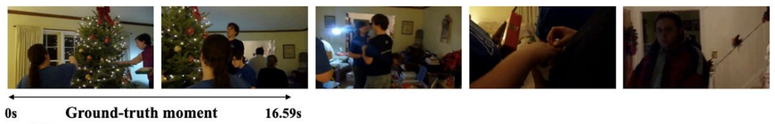
Can Video Large Language Models Comprehend Language in Videos?

Minjoon Jung[1,2], Xiao Junbin[1], Byoung-Tak Zhang[2], Angela Yao[1]
[1]National University of Singapore, [2]Seoul National University

Introduction



- Video Large Language Models (Video-LLMs) have shown great capabilities in video temporal understanding.
- However, such capabilities have not been thoroughly verified to be robust and trustable yet. Specifically, is their performance truly grounded in video temporal comprehension?
- In this study, we explore the consistency of Video-LLMs - a critical indicator for robust and trustworthy video temporal comprehension.

Evaluation

Query: A young girl is outside raking leaves out of the backyard.

Aligned & Misaligned Sentences

Replace Words
A young girl is outside gathering leaves from the backyard. → A young boy is inside scattering leaves in the living room.

Active to Passive
Outside, a young girl is raking leaves out of the backyard. → Inside, a young girl is scattering leaves around the house.

Word Order Changes
Leaves are being raked out of the backyard by a young girl outside. → Leaves are being scattered in the backyard by a young girl inside.

Compositional Information

A young girl is outside | raking leaves | out of the backyard.
Decompose
Is there a young girl in the backyard?
Is a young girl raking leaves?
Is there an old man inside the house?
Is an old man scattering leaves?

Consistent Temporal Grounding

When does the event "A young girl is outside raking a leaves out of the backyard." occur in the video? → VLLM → ***0 to 23 seconds***

When does the event "Leaves are being raked out of the backyard by a young girl outside" occur in the video? → VLLM → ***3 to 30 seconds***

When does the event "Outside, a young girl is raking leaves out of the backyard." occur in the video? → VLLM → ***10 to 40 seconds***

Self-answer Verification

*Does the event "A young girl is outside raking a leaves out of the backyard." occur in **0 to 23 seconds** in the video?* → VLLM → **Yes**

*Does the event "Inside, a young girl is scattering leaves around the house" occur in **0 to 23 seconds** in the video?* → VLLM → **No**

Compositional Understanding

*Is there a young girl in the backyard in **0 to 23 seconds** in the video?* → VLLM → **Yes**

*Is a young girl raking leaves in **0 to 23 seconds** in the video?* → VLLM → **Yes**

*Is an old man scattering leaves in **0 to 23 seconds** in the video?* → VLLM → **No**

Experiment Results

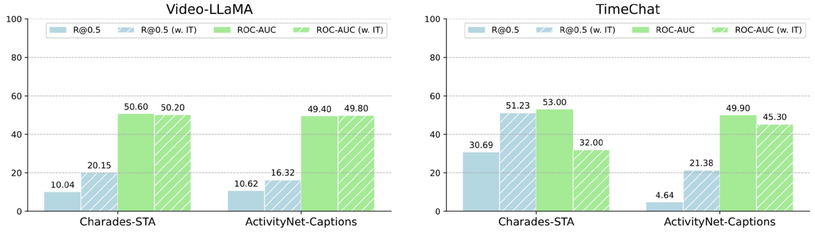
- Consistent Temporal Grounding

Methods	Charades-STA				ActivityNet-Captions			
	R@1,0.5	R@1,0.7	R_{con}@1,0.5	R_{con}@1,0.7	R@1,0.5	R@1,0.7	R_{con}@1,0.5	R_{con}@1,0.7
Video-LLaMA [16]	10.04	2.55	53.52	48.36	10.62	4.01	56.51	54.53
Video-ChatGPT [13]	14.43	7.64	89.22	87.90	6.68	2.95	64.56	63.86
TimeChat [3]	30.69	13.15	80.49	64.06	4.64	2.04	64.14	58.59
VTimeLLM [4]	27.72	11.88	83.16	80.61	31.43	17.16	83.30	78.82

- Self-answer Verification and Compositional Understanding

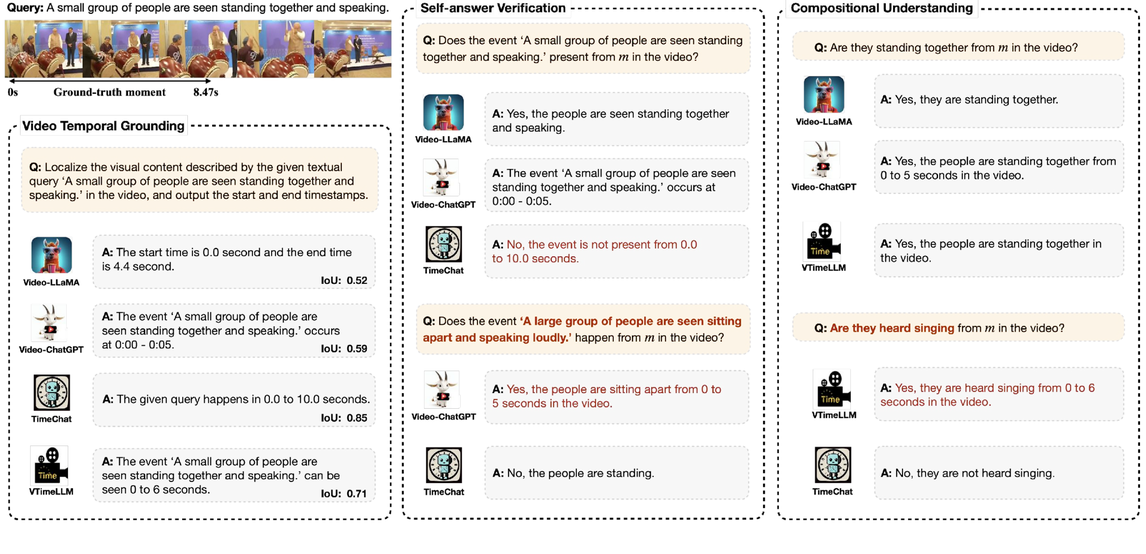
Methods	Charades-STA		ActivityNet-Captions	
	Self-answer Verification	Compositional Understanding	Self-answer Verification	Compositional Understanding
Random	50.0	50.0	50.0	50.0
Video-LLaMA [16]	50.6	49.7	49.4	53.4
Video-ChatGPT [13]	52.0	51.8	51.0	49.4
TimeChat [3]	53.0	55.7	49.9	51.9
VTimeLLM [4]	52.0	51.7	50.8	52.4

- The results of fine-tuned Video-LLMs



Qualitative Results

- Visualization of qualitative examples with four Video-LLMs.The **red text** represents using the misaligned sentence in the question and the Video-LLMs' inconsistent answer. These inconsistencies highlight significant gaps in the model's ability to provide reliable predictions.



Conclusion

- In this study, we investigate whether the temporal predictions of existing Video-LLMs are truly based on video language comprehension.
- Our results reveal that most Video-LLMs exhibit inconsistent answers and fine-tuning with target datasets does not improve consistency.
- Extended version with more comprehensive analyses and our solution for improvements:

Link

- Please scan the QR codes for more details.

Paper | Paper (Extended ver.) | Author