On the Consistency of Video Large Language Models in Temporal Comprehension


Minjoon Jung
Seoul National University


Junbin Xiao
National University of Singapore


Byoung Tak Zhang
Seoul National University


Angela Yao
National University of Singapore

CVPR Nashville JUNE 11-15, 2025

VERITAS LUX MEA

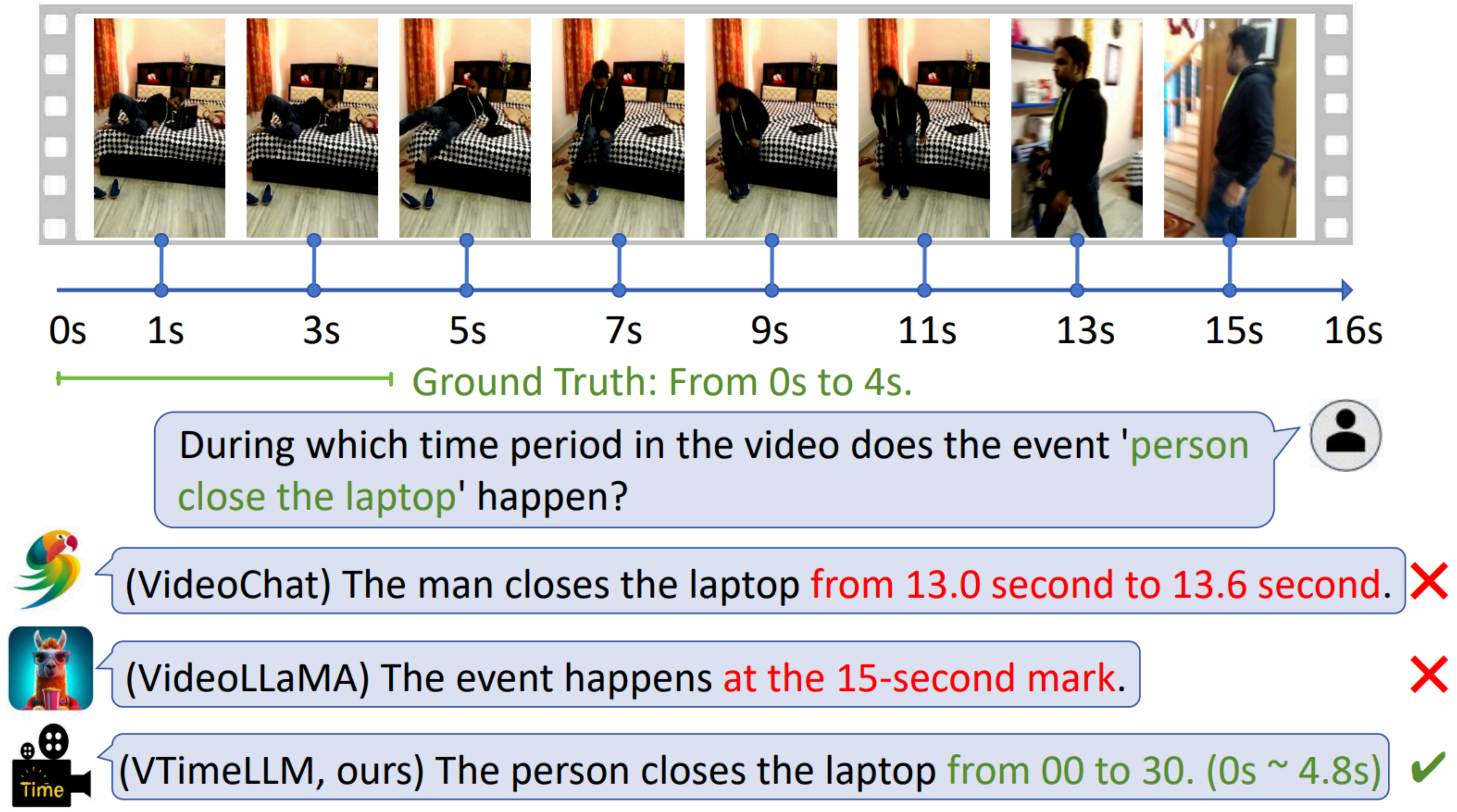
NUS National University of Singapore

Introduction

- Video large language models (Video-LLMs) can temporally ground language queries and retrieve video moments.

Question Answering (Coarse-grained)

Temporal Grounding / Dense Captioning (Fine-grained)

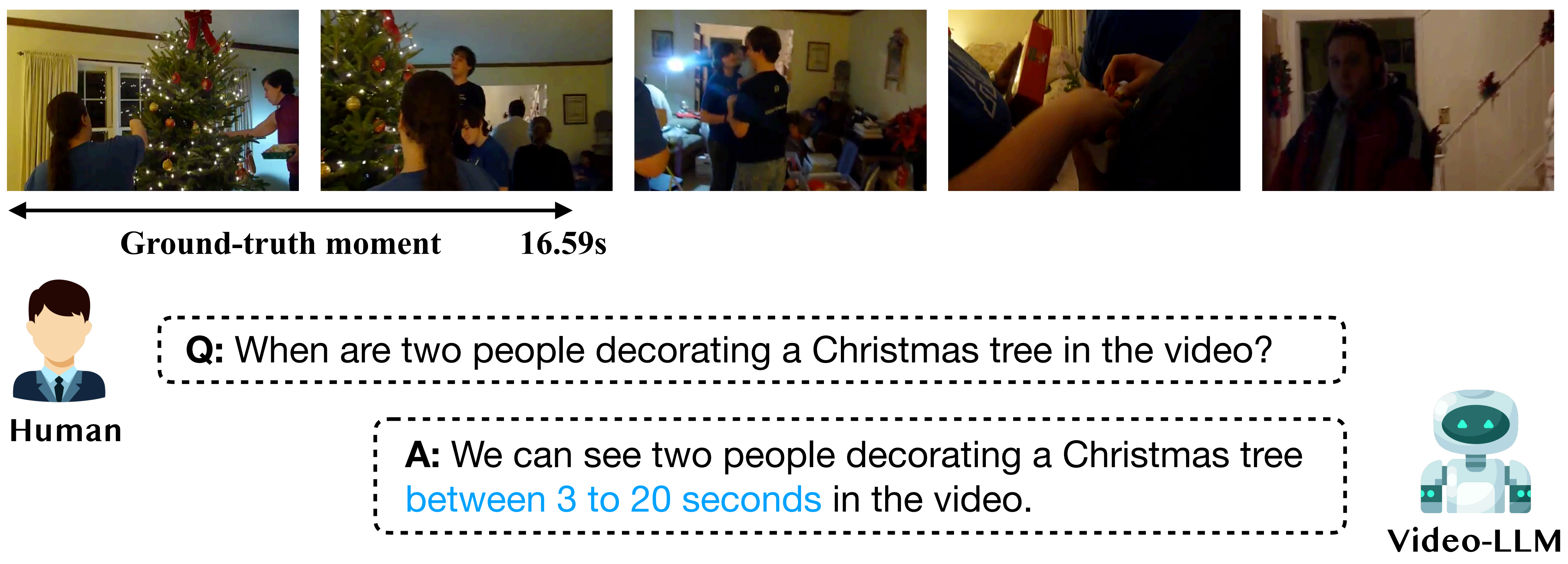


Beyond naive question-answering tasks, existing Video-LLMs are able to generate and identify fine-grained details, such as timestamps or segment-level captions in videos.

Introduction

- Yet, such temporal comprehension capabilities in Video-LLMs are neither well-studied nor understood. Specifically, is their performance truly grounded in video comprehension, or is it due to other spurious factors and correlations?

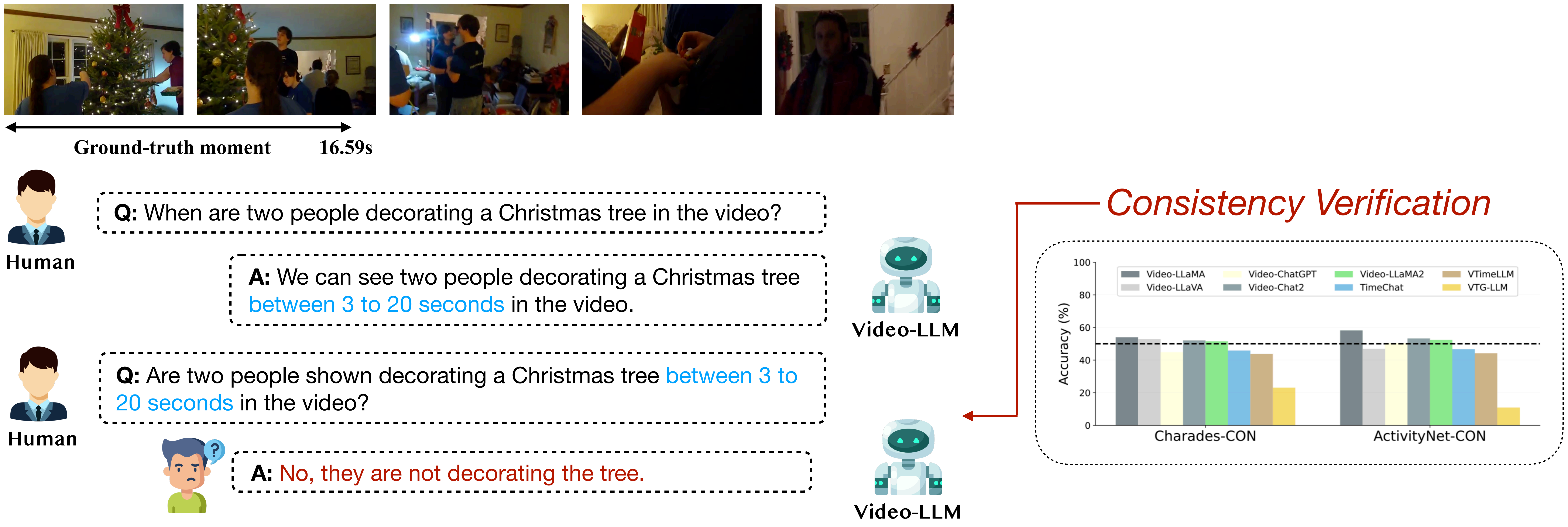
Video temporal grounding →



Given video and its query, the model grasps the specific moment i.e., start and end timestamps.

Introduction

- So we conduct a study on prediction consistency – a key indicator for robustness and trustworthiness of temporal grounding.
- We reveal that most Video-LLMs struggle to reliably confirm their initial moment predictions, achieving a near chance-level consistency (50%).



Video-LLMs are often inconsistent when asked to check their own video moment predictions, raising concern about their capabilities in faithful temporal reasoning.

Consistency Evaluation

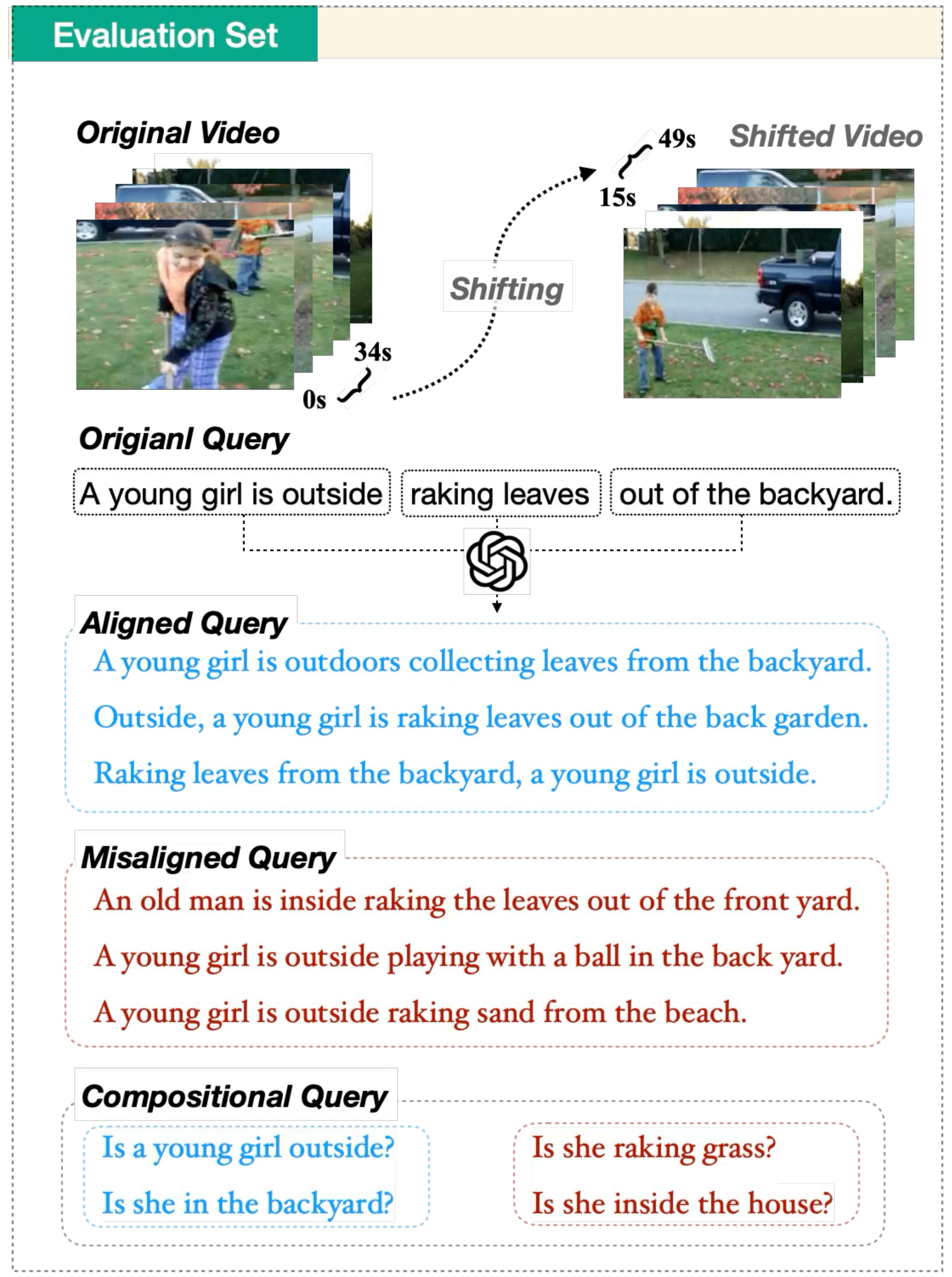
- To study the consistency of Video-LLMs in temporal comprehension, we first define two qualities.

  - **Grounding Consistency:** A model that links a language query to visual content in a specific temporal moment should remain robust to rephrasing queries and temporal shifts in the visual content.

  - **Verification Consistency:** If a model can identify a specific moment, its understanding should be bidirectional. As such, the model should be able to confirm that the said event and its components did occur within the predicted moment.

Consistency Evaluation

- To evaluate defined qualities, we construct new evaluation sets, Charades-CON and ActivityNet-CON, extending two popular video temporal grounding datasets: Charades-STA and ActivityNet-Captions.

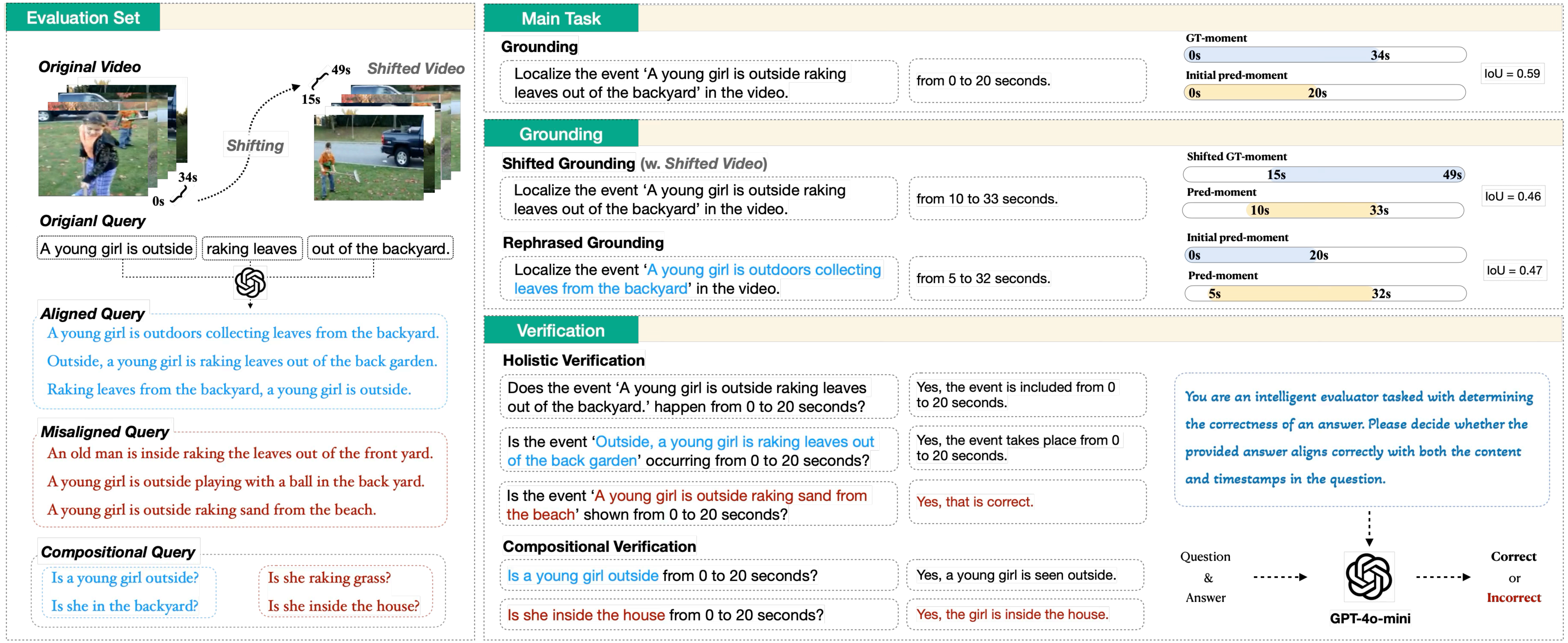


1. For grounding consistency, we generate 1) rephrased queries from the original queries using a powerful LLM (i.e., GPT-4o-mini) and 2) shifted videos through randomly shifting the ground-truth moment to different video segment.

2. For verification consistency, we design three different types of queries: aligned, misaligned, and compositional query. Queries are labeled as “aligned” or “misaligned” depending on whether they have the same or different meanings from the original query, respectively. Compositional queries partially overlap with the original by retaining some information, and we also generate them in both aligned and misaligned forms.

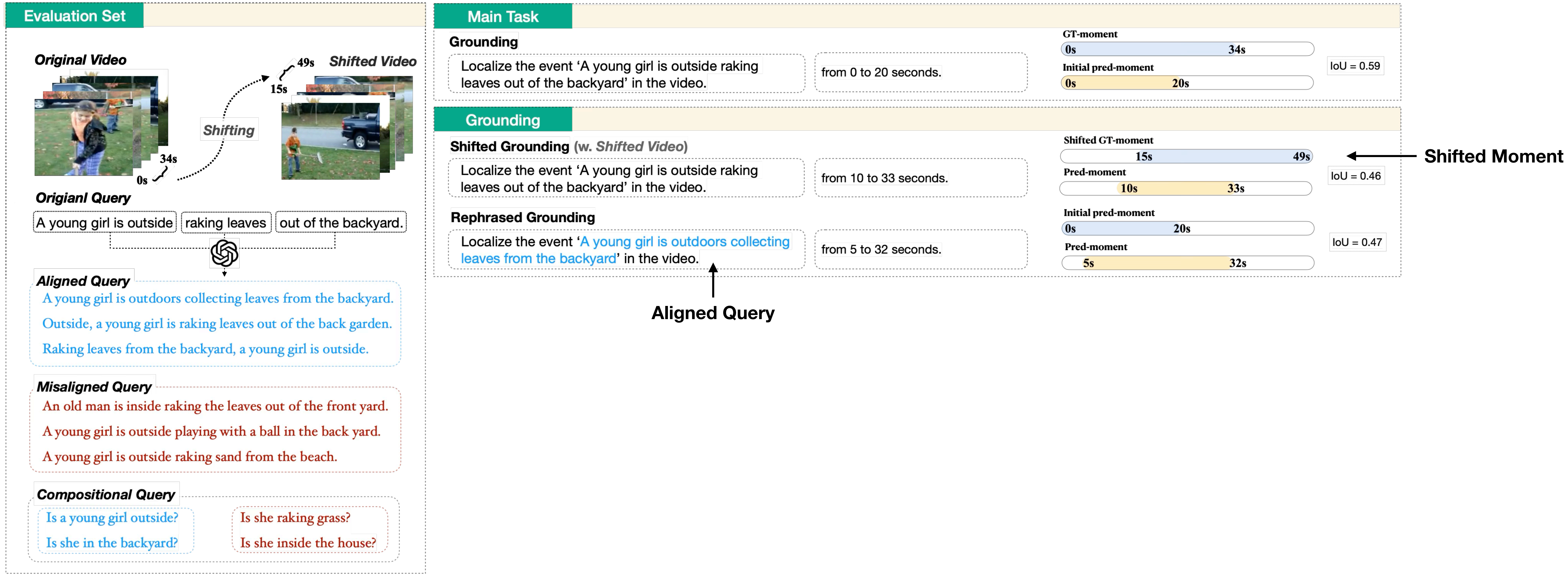
Consistency Evaluation

- With the constructed test sets, we design four probes: rephrased grounding, shifted grounding, holistic verification, and compositional verification.



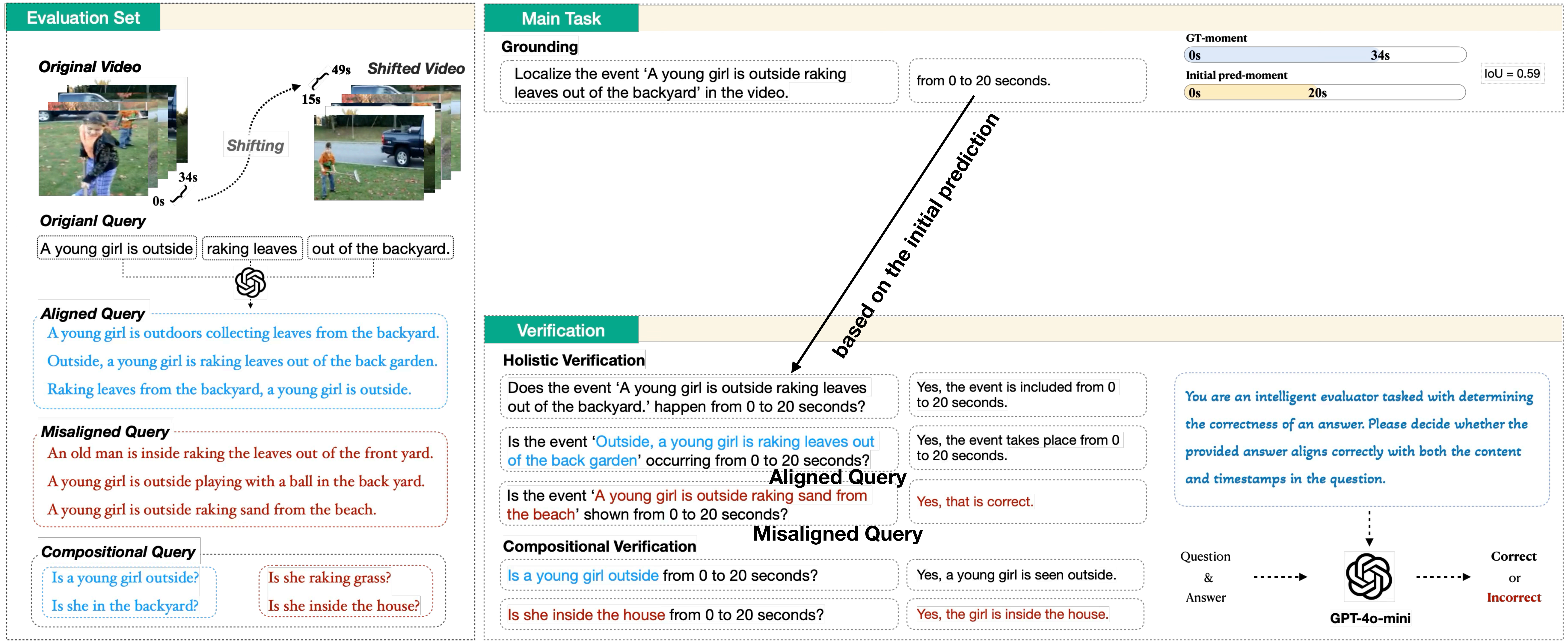
Consistency Evaluation

- Shifted grounding confirms whether a model adapt to visual variant.
- Rephrased grounding compares model predictions between the original query and its aligned variations.

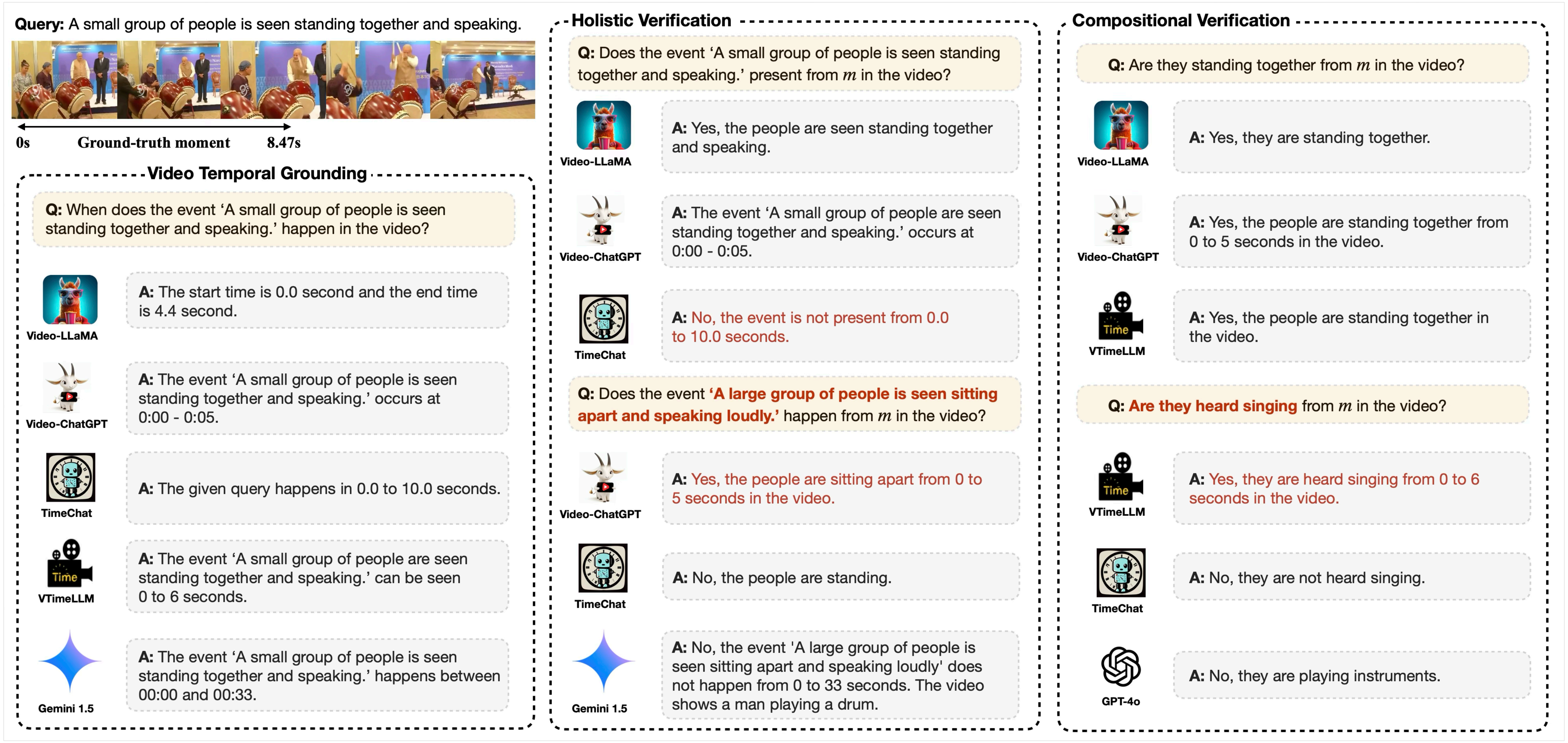


Consistency Evaluation

- Verification probes confirm whether the model can confirm the presence of the query within the moment prediction, aligning with its initial prediction.



Visualization



Experiments

- We include eight open-source models and two closed-source models in our evaluation.
- Open-source models can be categorized into two folds: general-purpose and time-aware models. Time-aware models are specifically designed to answer questions requiring precise timestamps within videos.
- We focus on correctly grounded queries, as evaluating consistency on incorrect predictions is less meaningful. So we report both the **absolute consistency score** (consistent predictions across the whole test set) and the **relative consistency score** (normalized within accurately grounded predictions).

Experiments

Method	LLM	# Frames	Charades-CON					ActivityNet-CON				
			Ground	R-Ground	S-Ground	H-Verify	C-Verify	Ground	R-Ground	S-Ground	H-Verify	C-Verify
Open-source (general-purpose)												
VideoChat2 [16]	Vicuna-7B	16	7.2	5.4 (74.5)	1.0 (13.7)	3.8 (52.1)	3.6 (50.0)	10.5	8.7 (82.8)	0.6 (6.0)	5.6 (53.3)	5.4 (51.9)
Video-LLaVA [19]	Vicuna-7B	8	9.4	7.6 (80.8)	2.8 (30.3)	5.0 (52.8)	4.7 (50.0)	13.4	10.0 (74.5)	3.1 (23.0)	6.3 (46.9)	7.0 (52.3)
Video-LLaMA [40]	Vicuna-7B	8	14.2	10.6 (74.9)	5.3 (37.6)	7.5 (53.3)	7.3 (51.7)	12.8	8.5 (66.8)	7.2 (56.8)	7.3 (57.5)	7.5 (58.9)
Video-ChatGPT [23]	Vicuna-7B	100	14.4	12.8 (89.2)	1.3 (8.8)	6.5 (44.8)	7.2 (50.0)	3.3	2.8 (84.0)	0.1 (4.1)	1.7 (50.4)	1.6 (49.2)
Video-LLaMA2 [4]	Mistral-7B	8	20.0	16.8 (83.8)	3.8 (19.0)	10.3 (51.5)	10.6 (52.9)	10.4	8.2 (78.6)	1.5 (14.8)	5.4 (52.4)	5.7 (54.7)
Open-source (time-aware)												
VTG-LLM [8]	Llama2-7B	96	26.0	16.1 (62.1)	**8.3** (32.0)	6.0 (23.1)	10.0 (38.4)	6.8	5.3 (78.0)	0.2 (3.0)	0.7 (10.9)	1.7 (24.9)
VTimeLLM [9]	Vicuna-7B	100	27.3	22.7 (83.2)	7.3 (26.9)	11.9 (43.7)	13.6 (49.8)	**31.9**	**26.4** (82.7)	**8.9** (27.9)	**14.1** (44.2)	**15.8** (49.5)
TimeChat [30]	Llama2-7B	96	**30.5**	**25.0** (82.1)	5.6 (18.5)	**14.0** (45.9)	**15.6** (51.2)	4.6	2.9 (64.1)	1.0 (21.2)	2.1 (46.7)	2.4 (52.2)
Closed-source												
GPT-4o [1]	-	10	28.5	21.2 (74.3)	9.3 (32.8)	17.8 (62.4)	20.3 (71.3)	26.8	18.1 (67.5)	10.4 (38.8)	16.5 (61.7)	18.4 (68.8)
Gemini 1.5 Flash [29]	-	1 fps	34.6	29.7 (85.7)	24.8 (71.7)	22.8 (65.8)	24.5 (70.8)	37.8	30.8 (81.4)	24.8 (65.6)	22.4 (59.3)	26.8 (70.8)

Table 1. Consistency evaluation of Video-LLMs and closed-source models. The *time-aware* models are specifically designed to grasp temporal moments. For each model, we specify the language model backbone (LLM) and the number of input frames (# Frames) used. Relative consistency scores are in brackets. Video-LLMs often struggle to consistently respond to the probes, revealing their deficiencies in reliable video temporal understanding. In contrast, the closed-source models demonstrate relatively superior consistency across all probes.

Experiments: Grounding

Method	LLM	# Frames	Charades-CON Ground	Charades-CON R-Ground	Charades-CON S-Ground	Charades-CON H-Verify	Charades-CON C-Verify	ActivityNet-CON Ground	ActivityNet-CON R-Ground	ActivityNet-CON S-Ground	ActivityNet-CON H-Verify	ActivityNet-CON C-Verify
Open-source (general-purpose)												
VideoChat2 [16]	Vicuna-7B	16	7.2	5.4 (74.5)	1.0 (13.7)	3.8 (52.1)	3.6 (50.0)	10.5	8.7 (82.8)	0.6 (6.0)	5.6 (53.3)	5.4 (51.9)
Video-LLaVA [19]	Vicuna-7B	8	9.4	7.6 (80.8)	2.8 (30.3)	5.0 (52.8)	4.7 (50.0)	13.4	10.0 (74.5)	3.1 (23.0)	6.3 (46.9)	7.0 (52.3)
Video-LLaMA [40]	Vicuna-7B	8	14.2	10.6 (74.9)	5.3 (37.6)	7.5 (53.3)	7.3 (51.7)	12.8	8.5 (66.8)	7.2 (56.8)	7.3 (57.5)	7.5 (58.9)
Video-ChatGPT [23]	Vicuna-7B	100	14.4	12.8 (89.2)	1.3 (8.8)	6.5 (44.8)	7.2 (50.0)	3.3	2.8 (84.0)	0.1 (4.1)	1.7 (50.4)	1.6 (49.2)
Video-LLaMA2 [4]	Mistral-7B	8	20.0	16.8 (83.8)	3.8 (19.0)	10.3 (51.5)	10.6 (52.9)	10.4	8.2 (78.6)	1.5 (14.8)	5.4 (52.4)	5.7 (54.7)
Open-source (time-aware)												
VTG-LLM [8]	Llama2-7B	96	26.0	16.1 (62.1)	**8.3** (32.0)	6.0 (23.1)	10.0 (38.4)	6.8	5.3 (78.0)	0.2 (3.0)	0.7 (10.9)	1.7 (24.9)
VTimeLLM [9]	Vicuna-7B	100	27.3	22.7 (83.2)	7.3 (26.9)	11.9 (43.7)	13.6 (49.8)	**31.9**	**26.4** (82.7)	**8.9** (27.9)	**14.1** (44.2)	**15.8** (49.5)
TimeChat [30]	Llama2-7B	96	**30.5**	**25.0** (82.1)	5.6 (18.5)	**14.0** (45.9)	**15.6** (51.2)	4.6	2.9 (64.1)	1.0 (21.2)	2.1 (46.7)	2.4 (52.2)
Closed-source												
GPT-4o [1]	-	10	28.5	21.2 (74.3)	9.3 (32.8)	17.8 (62.4)	20.3 (71.3)	26.8	18.1 (67.5)	10.4 (38.8)	16.5 (61.7)	18.4 (68.8)
Gemini 1.5 Flash [29]	-	1 fps	34.6	29.7 (85.7)	24.8 (71.7)	22.8 (65.8)	24.5 (70.8)	37.8	30.8 (81.4)	24.8 (65.6)	22.4 (59.3)	26.8 (70.8)

Table 1. Consistency evaluation of Video-LLMs and closed-source models. The *time-aware* models are specifically designed to grasp temporal moments. For each model, we specify the language model backbone (LLM) and the number of input frames (# Frames) used. Relative consistency scores are in brackets. Video-LLMs often struggle to consistently respond to the probes, revealing their deficiencies in reliable video temporal understanding. In contrast, the closed-source models demonstrate relatively superior consistency across all probes.

- *The time-aware models demonstrate superior grounding performance compared to general-purpose models.*
- *However, some of them fall short on ActivityNet-CON.*
- *Closed-source models show superior performance on both datasets.*

Experiments: Grounding Consistency

Method	LLM	# Frames	Charades-CON					ActivityNet-CON				
			Ground	R-Ground	S-Ground	H-Verify	C-Verify	Ground	R-Ground	S-Ground	H-Verify	C-Verify
Open-source (general-purpose)												
VideoChat2 [16]	Vicuna-7B	16	7.2	5.4 (74.5)	1.0 (13.7)	3.8 (52.1)	3.6 (50.0)	10.5	8.7 (82.8)	0.6 (6.0)	5.6 (53.3)	5.4 (51.9)
Video-LLaVA [19]	Vicuna-7B	8	9.4	7.6 (80.8)	2.8 (30.3)	5.0 (52.8)	4.7 (50.0)	13.4	10.0 (74.5)	3.1 (23.0)	6.3 (46.9)	7.0 (52.3)
Video-LLaMA [40]	Vicuna-7B	8	14.2	10.6 (74.9)	5.3 (37.6)	7.5 (53.3)	7.3 (51.7)	12.8	8.5 (66.8)	7.2 (56.8)	7.3 (57.5)	7.5 (58.9)
Video-ChatGPT [23]	Vicuna-7B	100	14.4	12.8 (89.2)	1.3 (8.8)	6.5 (44.8)	7.2 (50.0)	3.3	2.8 (84.0)	0.1 (4.1)	1.7 (50.4)	1.6 (49.2)
Video-LLaMA2 [4]	Mistral-7B	8	20.0	16.8 (83.8)	3.8 (19.0)	10.3 (51.5)	10.6 (52.9)	10.4	8.2 (78.6)	1.5 (14.8)	5.4 (52.4)	5.7 (54.7)
Open-source (time-aware)												
VTG-LLM [8]	Llama2-7B	96	26.0	16.1 (62.1)	**8.3** (32.0)	6.0 (23.1)	10.0 (38.4)	6.8	5.3 (78.0)	0.2 (3.0)	0.7 (10.9)	1.7 (24.9)
VTimeLLM [9]	Vicuna-7B	100	27.3	22.7 (83.2)	7.3 (26.9)	11.9 (43.7)	13.6 (49.8)	**31.9**	**26.4** (82.7)	**8.9** (27.9)	**14.1** (44.2)	**15.8** (49.5)
TimeChat [30]	Llama2-7B	96	**30.5**	**25.0** (82.1)	5.6 (18.5)	**14.0** (45.9)	**15.6** (51.2)	4.6	2.9 (64.1)	1.0 (21.2)	2.1 (46.7)	2.4 (52.2)
Closed-source												
GPT-4o [1]	-	10	28.5	21.2 (74.3)	9.3 (32.8)	17.8 (62.4)	20.3 (71.3)	26.8	18.1 (67.5)	10.4 (38.8)	16.5 (61.7)	18.4 (68.8)
Gemini 1.5 Flash [29]	-	1 fps	34.6	29.7 (85.7)	24.8 (71.7)	22.8 (65.8)	24.5 (70.8)	37.8	30.8 (81.4)	24.8 (65.6)	22.4 (59.3)	26.8 (70.8)

Table 1. Consistency evaluation of Video-LLMs and closed-source models. The *time-aware* models are specifically designed to grasp temporal moments. For each model, we specify the language model backbone (LLM) and the number of input frames (# Frames) used. Relative consistency scores are in brackets. Video-LLMs often struggle to consistently respond to the probes, revealing their deficiencies in reliable video temporal understanding. In contrast, the closed-source models demonstrate relatively superior consistency across all probes.

- All models behave relatively well in this probe, with a relative consistency score exceeding 60%, likely due to the use of LLMs for language comprehension.

- All open-source models struggle to adapt their predictions to shifted moments, consistently showing lower scores than in rephrased grounding. This suggests that they are less sensitive to visual variants than language variants.

Experiments: Verification Consistency

Method	LLM	# Frames	Charades-CON					ActivityNet-CON				
			Ground	R-Ground	S-Ground	H-Verify	C-Verify	Ground	R-Ground	S-Ground	H-Verify	C-Verify
Open-source (general-purpose)												
VideoChat2 [16]	Vicuna-7B	16	7.2	5.4 (74.5)	1.0 (13.7)	3.8 (52.1)	3.6 (50.0)	10.5	8.7 (82.8)	0.6 (6.0)	5.6 (53.3)	5.4 (51.9)
Video-LLaVA [19]	Vicuna-7B	8	9.4	7.6 (80.8)	2.8 (30.3)	5.0 (52.8)	4.7 (50.0)	13.4	10.0 (74.5)	3.1 (23.0)	6.3 (46.9)	7.0 (52.3)
Video-LLaMA [40]	Vicuna-7B	8	14.2	10.6 (74.9)	5.3 (37.6)	7.5 (53.3)	7.3 (51.7)	12.8	8.5 (66.8)	7.2 (56.8)	7.3 (57.5)	7.5 (58.9)
Video-ChatGPT [23]	Vicuna-7B	100	14.4	12.8 (89.2)	1.3 (8.8)	6.5 (44.8)	7.2 (50.0)	3.3	2.8 (84.0)	0.1 (4.1)	1.7 (50.4)	1.6 (49.2)
Video-LLaMA2 [4]	Mistral-7B	8	20.0	16.8 (83.8)	3.8 (19.0)	10.3 (51.5)	10.6 (52.9)	10.4	8.2 (78.6)	1.5 (14.8)	5.4 (52.4)	5.7 (54.7)
Open-source (time-aware)												
VTG-LLM [8]	Llama2-7B	96	26.0	16.1 (62.1)	**8.3** (32.0)	6.0 (23.1)	10.0 (38.4)	6.8	5.3 (78.0)	0.2 (3.0)	0.7 (10.9)	1.7 (24.9)
VTimeLLM [9]	Vicuna-7B	100	27.3	22.7 (83.2)	7.3 (26.9)	11.9 (43.7)	13.6 (49.8)	**31.9**	**26.4** (82.7)	**8.9** (27.9)	**14.1** (44.2)	**15.8** (49.5)
TimeChat [30]	Llama2-7B	96	**30.5**	**25.0** (82.1)	5.6 (18.5)	**14.0** (45.9)	**15.6** (51.2)	4.6	2.9 (64.1)	1.0 (21.2)	2.1 (46.7)	2.4 (52.2)
Closed-source												
GPT-4o [1]	-	10	28.5	21.2 (74.3)	9.3 (32.8)	17.8 (62.4)	20.3 (71.3)	26.8	18.1 (67.5)	10.4 (38.8)	16.5 (61.7)	18.4 (68.8)
Gemini 1.5 Flash [29]	-	1 fps	34.6	29.7 (85.7)	24.8 (71.7)	22.8 (65.8)	24.5 (70.8)	37.8	30.8 (81.4)	24.8 (65.6)	22.4 (59.3)	26.8 (70.8)

Table 1. Consistency evaluation of Video-LLMs and closed-source models. The *time-aware* models are specifically designed to grasp temporal moments. For each model, we specify the language model backbone (LLM) and the number of input frames (# Frames) used. Relative consistency scores are in brackets. Video-LLMs often struggle to consistently respond to the probes, revealing their deficiencies in reliable video temporal understanding. In contrast, the closed-source models demonstrate relatively superior consistency across all probes.

- *All open-source models demonstrate near chance-level verification consistency (with relative consistency scores around 50%).*
- *Specifically, most models are only able to respond consistently for about half of their correct moment predictions, indicating significant room for improvement in their verification consistency.*

Discussion and Solution

- Overall results suggest that <u>Video-LLMs exhibit almost the same degree of inconsistencies across the datasets</u>, regardless of their particular designs.
- Here, we investigate how two common solutions: prompting and instruction tuning with target datasets affect the model's consistency.

Prompting

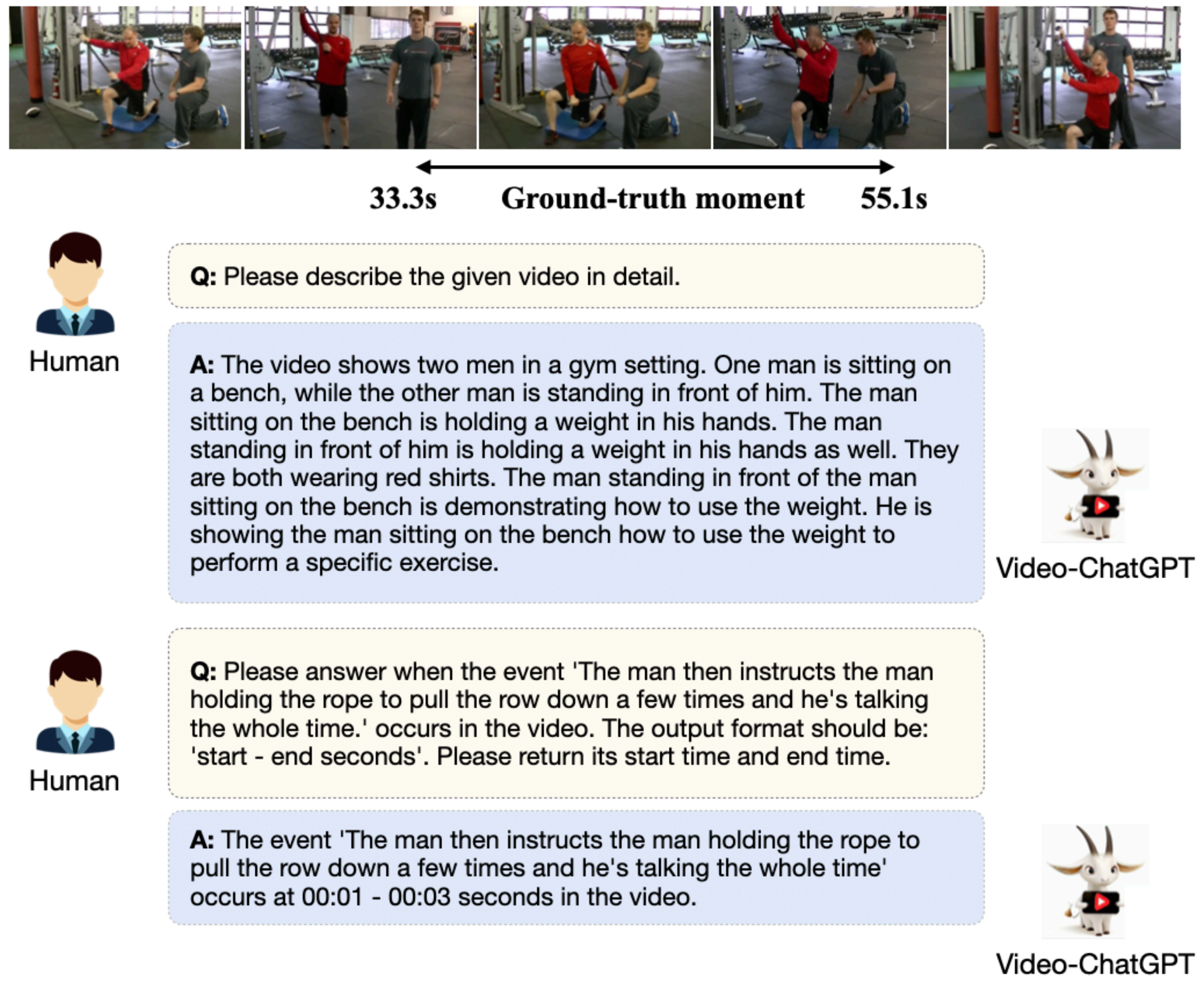
- We devise two prompting methods.
  - The first method, inspired by Chain-of-Thought (CoT), provides step-by-step guidance to solve the target tasks. We describe each probe and offer instructions for tackling it sequentially.
  - The second method is description prompting (Desc) that prompts the model to describe the video content before attempting the target tasks.

Grounding / Rephrased Grounding / Shifted Grounding

Your task is to predict the start and end times of an action or event described by a query sentence based on the visual content of the video. Use Chain-of-Thought reasoning to break down the query, analyze key moments, and accurately identify the time range where the action occurs.

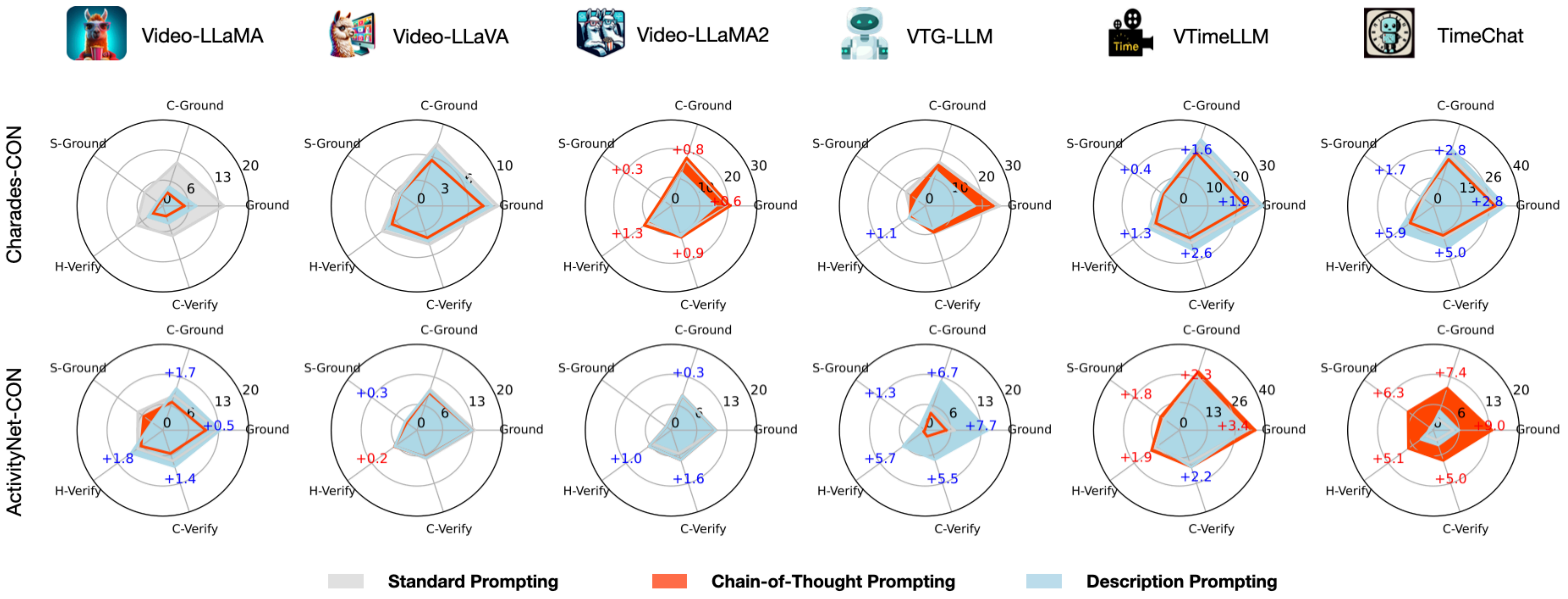
Chain-of-Thought Reasoning:
1. **Step 1: Parse the Query**: Break down the query sentence to understand the key action or event that you need to locate.
2. **Step 2: Analyze the Video Features**: Examine the sequence of video frames to detect patterns that match the key action described in the query.
3. **Step 3: Identify the Temporal Boundaries**: Use temporal reasoning to find the start and end frames of the action based on the video features.
4. **Step 4: Predict Start and End Times**: Map the identified frames to timestamps in the video, making sure the start and end times align with the query.
5. **Step 5: Verify the Answer**: Check if the predicted time range accurately captures the action described in the query.

Chain-of-Thought prompting



Description prompting

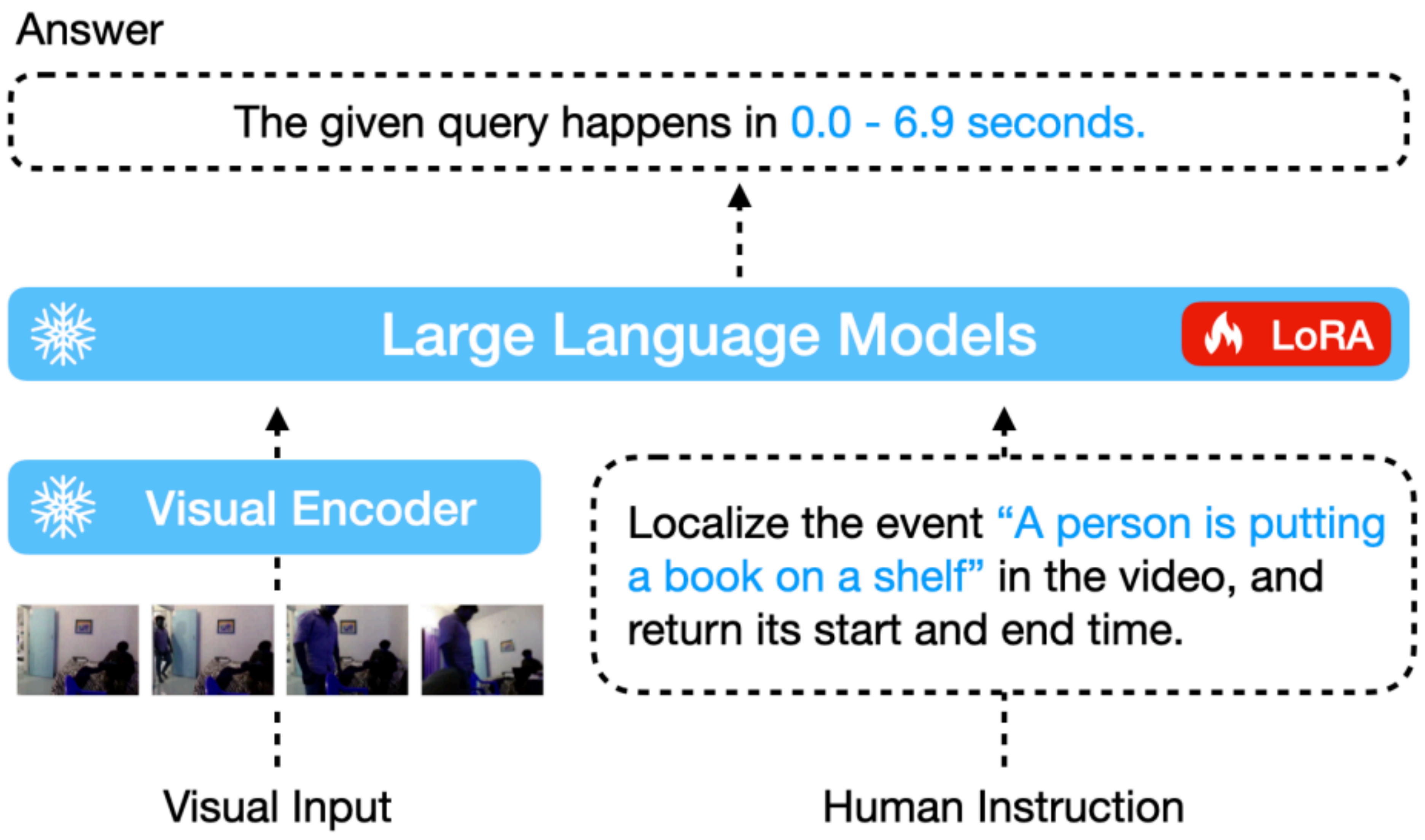
Prompting



- *CoT demonstrates over a 5% improvement across all aspects of TimeChat on ActivityNet-CON.*
- *However, none of the prompting methods improve Video-LLaMA and VideoLLaVA on Charades-CON; in fact, they even degrade the performance.*
- *Overall, <u>the improvements are often unstable</u>, underscoring the limitations of prompting alone in enhancing temporal comprehension.*

Fine-Tuning

- For instruction tuning, we select one general-purpose model, Video-LLaMA, and one time-aware model, TimeChat. Then, following their official configurations, we conduct instruction tuning using pre-defined templates like “The given event occurs from {start} to {end} seconds.”



Fine-Tuning

Method	FT	Charades-CON					ActivityNet-CON				
		Ground	R-Ground	S-Ground	H-Verify	C-Verify	Ground	R-Ground	S-Ground	H-Verify	C-Verify
Video-LLaMA [40]	✗	14.2	10.6 (74.9)	5.3 (37.6)	7.5 (53.3)	7.3 (51.7)	12.8	8.5 (66.8)	7.2 (56.8)	7.3 (57.5)	7.5 (58.9)
Video-LLaMA (w. CoT)	✗	5.0	3.3 (66.6)	1.1 (22.2)	2.9 (58.7)	2.5 (50.0)	10.2	6.9 (68.0)	5.5 (54.4)	6.2 (61.5)	5.7 (56.2)
Video-LLaMA (w. Desc)	✗	7.7	4.7 (60.6)	1.5 (20.0)	4.3 (55.5)	4.2 (55.0)	13.3	10.2 (73.3)	3.1 (22.2)	9.1 (65.3)	8.9 (64.1)
Video-LLaMA (w. IT)	✓	45.1	32.0 (71.8)	8.9 (19.7)	24.8 (54.9)	23.4 (51.8)	20.6	16.4 (80.7)	6.5 (32.0)	11.7 (57.3)	11.1 (54.1)
TimeChat [30]	✗	30.5	25.0 (82.1)	5.6 (18.5)	14.0 (45.9)	15.6 (51.2)	4.6	2.9 (64.1)	1.0 (21.2)	2.1 (46.7)	2.4 (52.2)
TimeChat (w. CoT)	✗	28.7	22.8 (79.4)	7.1 (24.6)	13.5 (46.9)	14.4 (50.2)	13.6	10.3 (75.9)	7.3 (53.6)	7.2 (52.8)	7.4 (54.4)
TimeChat (w. Desc)	✗	33.3	27.8 (83.4)	7.3 (22.0)	19.9 (59.9)	20.6 (61.8)	5.9	4.7 (79.7)	0.7 (11.9)	3.6 (60.3)	3.6 (61.4)
TimeChat (w. IT)	✓	55.8	50.9 (91.3)	10.5 (18.9)	16.7 (30.0)	25.7 (46.2)	25.3	20.2 (80.4)	7.5 (29.9)	8.7 (34.5)	12.6 (49.9)

Table 2. Consistency evaluation of Video-LLMs with the proposed solutions. Performance gains are highlighted in blue, while decreased performances are shown in red. FT represents whether the model is fine-tuned. Compared to the previous solutions (*i.e.* CoT, Desc, and IT), VTune demonstrates significant improvements for both grounding and consistency across the models and datasets.

- Instruction Tuning consistently boosts the model's grounding performance but remains ineffective in improving consistency.
- Specifically, although IT significantly improves the grounding performance of TimeChat by 25.3% on CharadesCON, only a marginal improvement of 2.7% is observed in holistic verification due to its low relative consistency score.

Event Temporal Verification Tuning (VTune)

- We introduce Event Temporal Verification Tuning (VTune) that aims to verify their moment predictions.
- To achieve this, we intentionally changes the information (i.e., content and timestamps) in the query and ask the model to recognize the changes and correctly respond with reasoning.

Query: A person starts cooking.

12.4s ← Ground-truth moment → 27.0s

Instruction Tuning (IT)

Grounding Query

Q: Localize the visual content described by the given textual query 'A person starts cooking' in the video, and output the start and end timestamps.

A: The given query happens in 12.4 to 27.0 seconds.

Event Temporal Verification Tuning (VTune)

Event Verification Query

Q: Is the event '**The individual is preparing a dish.**' present from 12.4 to 27.0 seconds in the video?

A: **Yes**, we can see the event 'The individual is preparing a dish.' from 12.4 to 27.0 seconds.

Q: Does the event '**The person is eating dinner at a table.**' not happen from 12.4 to 27.0 seconds in the video?"

A: Yes, the event 'The person is eating dinner at a table.' does not happen from 12.4 to 27.0 seconds in the video. **We cannot see a person eating dinner; the individual is cooking.**

Temporal Verification Query

Q: Is the event 'A person starts cooking' present from **0 to 10 seconds** in the video?

A: **No**, we can see the event 'A person starts cooking' from **12.4 to 27.0 seconds**.

Event Temporal Verification Tuning (VTune)

- We demonstrate the effectiveness of VTune with extensive experiments.

Consistent performance improvements

Method	FT	Charades-CON					ActivityNet-CON				
		Ground	R-Ground	S-Ground	H-Verify	C-Verify	Ground	R-Ground	S-Ground	H-Verify	C-Verify
Video-LLaMA [40]	✗	14.2	10.6 (74.9)	5.3 (37.6)	7.5 (53.3)	7.3 (51.7)	12.8	8.5 (66.8)	7.2 (56.8)	7.3 (57.5)	7.5 (58.9)
Video-LLaMA (w. CoT)	✗	5.0	3.3 (66.6)	1.1 (22.2)	2.9 (58.7)	2.5 (50.0)	10.2	6.9 (68.0)	5.5 (54.4)	6.2 (61.5)	5.7 (56.2)
Video-LLaMA (w. Desc)	✗	7.7	4.7 (60.6)	1.5 (20.0)	4.3 (55.5)	4.2 (55.0)	13.3	10.2 (73.3)	3.1 (22.2)	9.1 (65.3)	8.9 (64.1)
Video-LLaMA (w. IT)	✓	45.1	32.0 (71.8)	8.9 (19.7)	24.8 (54.9)	23.4 (51.8)	20.6	16.4 (80.7)	6.5 (32.0)	11.7 (57.3)	11.1 (54.1)
Video-LLaMA **(w. VTune)**	✓	**54.4**	**38.2** (70.3)	**10.9** (20.0)	**30.7** (56.5)	**30.0** (55.2)	**33.0**	**24.7** (74.8)	**10.0** (30.2)	**20.2** (61.1)	**17.7** (53.7)
TimeChat [30]	✗	30.5	25.0 (82.1)	5.6 (18.5)	14.0 (45.9)	15.6 (51.2)	4.6	2.9 (64.1)	1.0 (21.2)	2.1 (46.7)	2.4 (52.2)
TimeChat (w. CoT)	✗	28.7	22.8 (79.4)	7.1 (24.6)	13.5 (46.9)	14.4 (50.2)	13.6	10.3 (75.9)	7.3 (53.6)	7.2 (52.8)	7.4 (54.4)
TimeChat (w. Desc)	✗	33.3	27.8 (83.4)	7.3 (22.0)	19.9 (59.9)	20.6 (61.8)	5.9	4.7 (79.7)	0.7 (11.9)	3.6 (60.3)	3.6 (61.4)
TimeChat (w. IT)	✓	55.8	50.9 (91.3)	10.5 (18.9)	16.7 (30.0)	25.7 (46.2)	25.3	20.2 (80.4)	7.5 (29.9)	8.7 (34.5)	12.6 (49.9)
TimeChat **(w. VTune)**	✓	**76.2**	**69.2** (90.8)	**36.2** (47.5)	**44.8** (58.8)	**42.4** (55.7)	**37.4**	**28.3** (75.6)	**10.6** (28.3)	**19.6** (52.3)	**19.3** (51.5)

Table 2. Consistency evaluation of Video-LLMs with the proposed solutions. Performance gains are highlighted in blue, while decreased performances are shown in red. FT represents whether the model is fine-tuned. Compared to the previous solutions (*i.e.* CoT, Desc, and IT), VTune demonstrates significant improvements for both grounding and consistency across the models and datasets.

Ablation studies on VTune

Method	G	E	T	Charades-CON					
				Ground	Ground (0.7)	R-Ground	S-Ground	H-Verify	C-Verify
Video-LLaMA	✓			45.1	23.3	32.0 (71.8)	8.9 (19.7)	24.8 (54.9)	23.4 (51.8)
	✓	✓		48.8	27.8	35.1 (71.9)	10.2 (20.8)	28.0 (57.4)	26.4 (54.1)
	✓	✓	✓	**54.4**	**36.6**	**38.2** (70.3)	**10.9** (20.0)	**30.7** (56.5)	**30.0** (55.2)
TimeChat	✓			55.8	30.2	50.9 (91.3)	10.5 (18.9)	16.7 (30.0)	25.7 (46.2)
	✓	✓		**76.2**	52.3	68.7 (90.1)	15.5 (20.4)	40.4 (53.0)	40.5 (53.1)
	✓	✓	✓	**76.2**	**58.8**	**69.2** (90.8)	**36.2** (47.5)	**44.8** (58.8)	**42.4** (55.7)

Table 4. Query effectiveness by type. G represents grounding queries, and E and T indicate event and temporal verification queries. Additionally, we report "R@1, IoU=0.7" in Ground (0.7).

State-of-the-art performances

Method	Charades-STA		ActivityNet-Captions	
	R@1, 0.5	R@1, 0.7	R@1, 0.5	R@1, 0.7
Task-Specific Models				
BM-DETR [11]	59.4	38.3	49.6	30.6
Mr.BLIP [2]	69.3	49.2	53.9	35.5
Video-LLMs				
HawkEye [32]	58.3	28.8	34.7	17.7
VTG-LLM [8]	57.2	33.4	-	-
Video-LLaMA [40]	35.0	18.6	25.2	14.4
Video-LLaMA **(w. VTune)**	37.1	20.1	34.3	19.1
TimeChat [30]	46.7	23.7	28.0	15.8
TimeChat **(w. VTune)**	**58.4**	**34.7**	**41.0**	**23.7**

Table 3. Fine-tuned performances on the original test sets. Task-specific models are designed to perform only a single target task. VTune effectively improves grounding performance in both models.

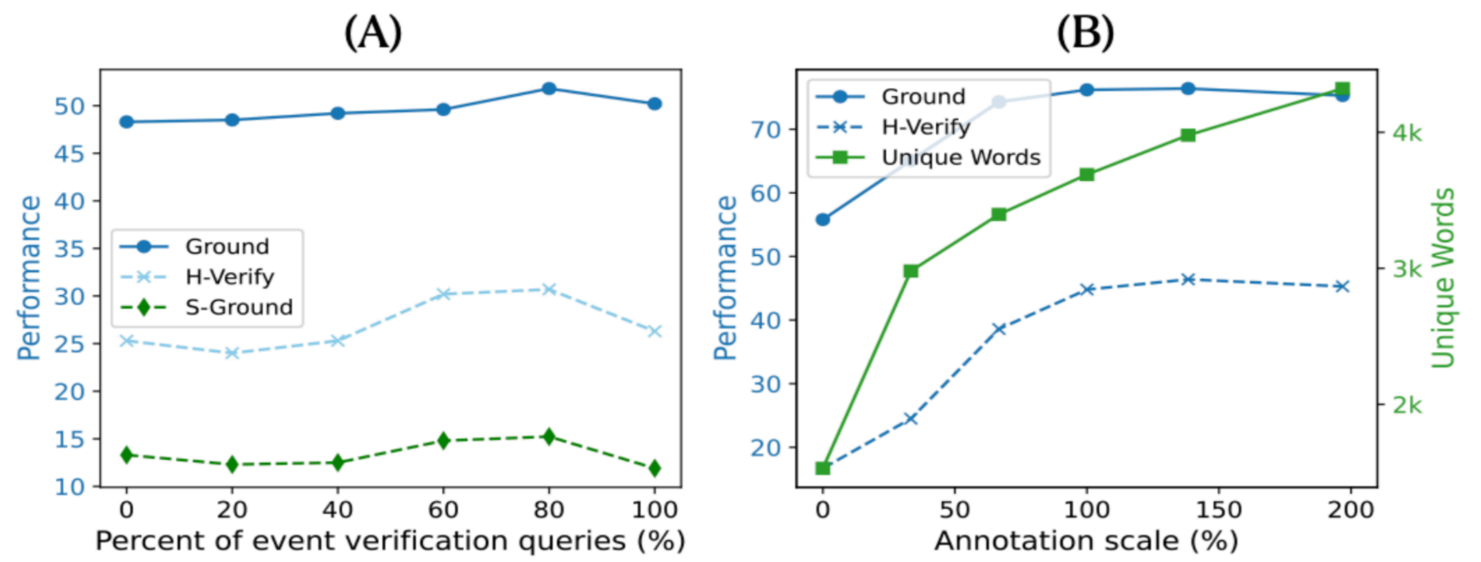


Figure 1. **Experiments on Charades-CON with TimeChat. (A)**: Performance evaluation across different ratios of event and temporal verification queries. The total queries is fixed at 12k, matching the # of grounding queries in the original dataset. **(B)**: Performance evaluation across varying verification query scales. 100% scale corresponds to the performance of VTune reported in the paper. The number of unique words is a proxy for dataset diversity.

Conclusions

- We have studied the consistency of Video-LLMs in temporal comprehension. We design evaluation datasets corresponding to a series of dedicated probes to analyze the consistency of Video-LLMs.
- Our findings show that most Video-LLMs exhibit inconsistent behaviors, unveiling their significant deficiencies in reliable video temporal comprehension. We further demonstrate the limitations of common prompting methods and instruction tuning with target datasets, revealing their unstable improvements.
- To this end, we propose VTune, which performs instruction tuning by explicitly encouraging consistency, demonstrating its significant improvement in both grounding and consistency.

Thank you for your attention!