🍅 TOMATO: Assessing Visual Temporal Reasoning Capabilities in Multimodal Foundation Models

Yale

Ziyao Shangguan*, Chuhan Li*, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, Arman Cohan

Ai2

Do Multimodal Foundation Models truly understand videos or just individual frames?

Motivation

Existing benchmarks do not mandate the video modality.

Existing Benchmarks

Video ID: 7966

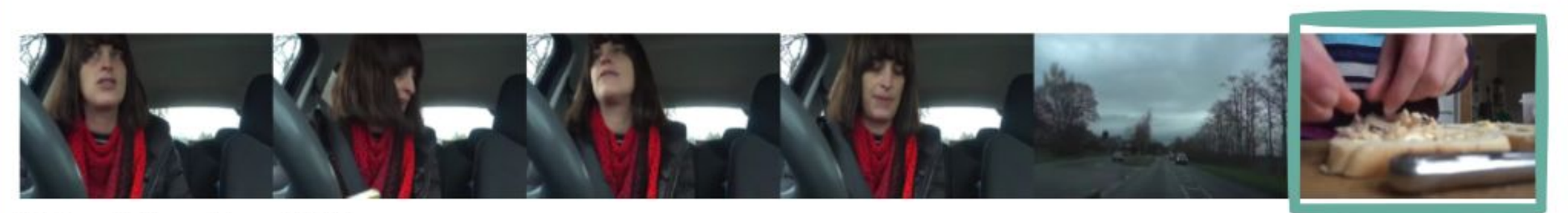
VITATECS (Li et al., 2023):
Which of the following best describes the content of the video?
A. Cheese is being spread B. Cheese is being sliced

Video ID: friends_s05e10_seg02_clip_11_3.52_37.14

MVBench (Li et al., 2024):
What is behind Monica when she is in the chair talking to Ross?
A. A television B. Joey's white dog sculpture
C. An American flag D. A Christmas tree E. A basket of laundry

Video ID: 1034736653

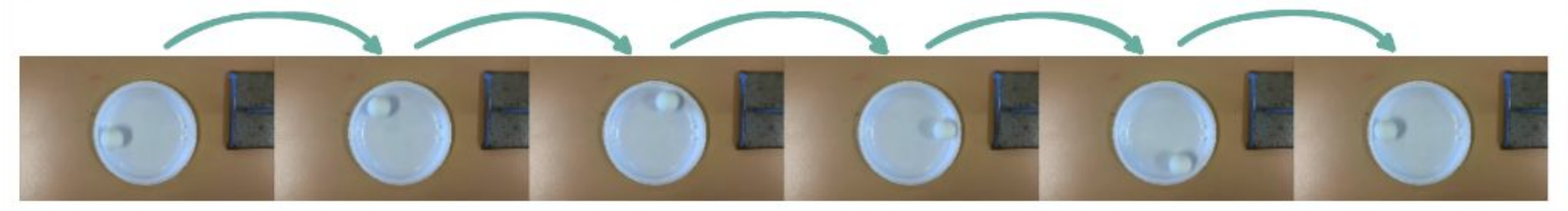
TempCompass (Liu et al., 2024):
What are the woman athletes doing?
A. Cycling B. Swimming C. Running D. Dancing



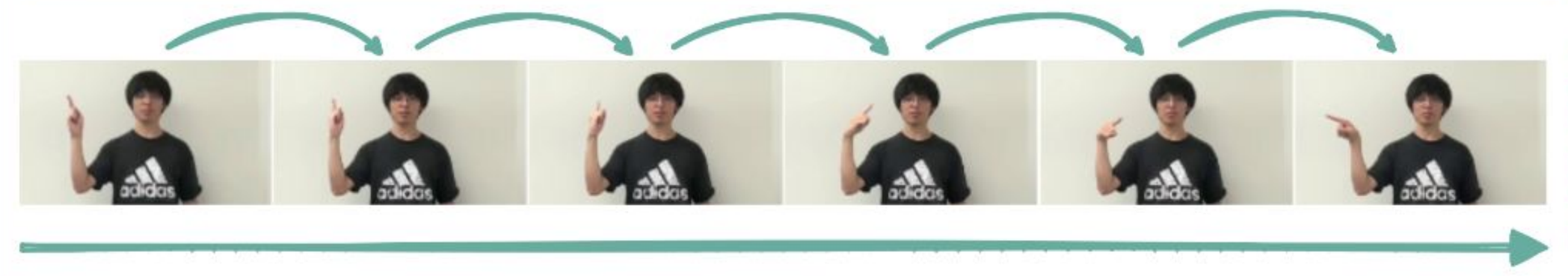
Video ID: qvh_val368

ReXTime (Chen et al., 2024):
Why is the woman preparing ingredients and utensils?
A. To make herself a sandwich B. To cook dinner for her family
C. To bake a cake D. To prepare a salad

🍅 TOMATO (Ours)



In which direction(s) does the Ping Pong ball rotate?
A. Clockwise throughout B. Counter-clockwise then clockwise
C. Counter-clockwise throughout D. Clockwise then counter-clockwise



What directional command is this person trying to convey?
A. Move to the left B. Move to the right
C. Move up D. Turn around

Our Benchmark: 🍅 TOMATO

- **1,484** human-annotated questions applied to **1,417** videos—including **805** self-recorded and generated videos, **398** YouTube videos, and **214** videos from existing datasets
- **6 reasoning types**: *rotation, direction, velocity & frequency, shape & trend, visual cues, action count*
- **3 scenarios**: *human-centric, real-world, simulated*

ICLR International Conference On Learning Representations

arXiv

Dataset

Benchmarking Principles

Our 3 principles aim to enforce visual temporal reasoning.

1. Multi-Frame Gain

$$\kappa = \frac{Acc(m \text{ frames}) - Acc(1 \text{ frame})}{Acc(1 \text{ frame}) + \epsilon}$$

*How much **better** does a model perform using multiple frames compared to one?*

A **higher κ** value on TOMATO indicates a necessity to reason across multiple frames, and the question cannot be accurately answered using a single frame.

	VITATECS			MVBench			TempCompass			ReXTime			TOMATO		
# Frames	1	16	κ ↑	1	16	κ ↑	1	16	κ ↑	1	16	κ ↑	1	16	κ ↑
Average	70.7	87.2	23.4	47.1	62.7	33.3	51.3	75.3	47.1	62.8	80.0	27.4	20.9	37.8	**81.0**
# Frames	1[H]	16	κ ↑	1[H]	16	κ ↑	1[H]	16	κ ↑	1[H]	16	κ ↑	1[H]	16	κ ↑
Average	86.0	87.2	1.5	58.8	62.7	6.7	64.0	75.3	17.6	86.1	80.0	-7.1	22.8	37.8	**66.3**

2. Frame-Order Sensitivity

$$\tau = \frac{Acc(m \text{ frames}) - Acc(\text{shuffled } m \text{ frames})}{Acc(\text{shuffled } m \text{ frames}) + \epsilon}$$

*How much does performance **increase** when we put shuffled frames back in the correct order?*

A **higher τ** value on TOMATO suggests a stronger reliance on the frames' correct order, and the question cannot be accurately answered using out-of-order frames.

	VITATECS			MVBench			TempCompass			ReXTime			TOMATO		
# Frames	16[S]	16	τ ↑	16[S]	16	τ ↑	16[S]	16	τ ↑	16[S]	16	τ ↑	16[S]	16	τ ↑
Average	84.5	87.2	3.2	59.3	62.7	5.8	61.5	75.3	22.3	79.7	80.0	0.3	28.5	37.8	**34.1**

3. Frame-Information Disparity

$$\rho = \frac{Acc(\text{handpicked 1 frame}) - Acc(\text{random-sampled 1 frame})}{Acc(\text{random-sampled 1 frame}) + \epsilon}$$

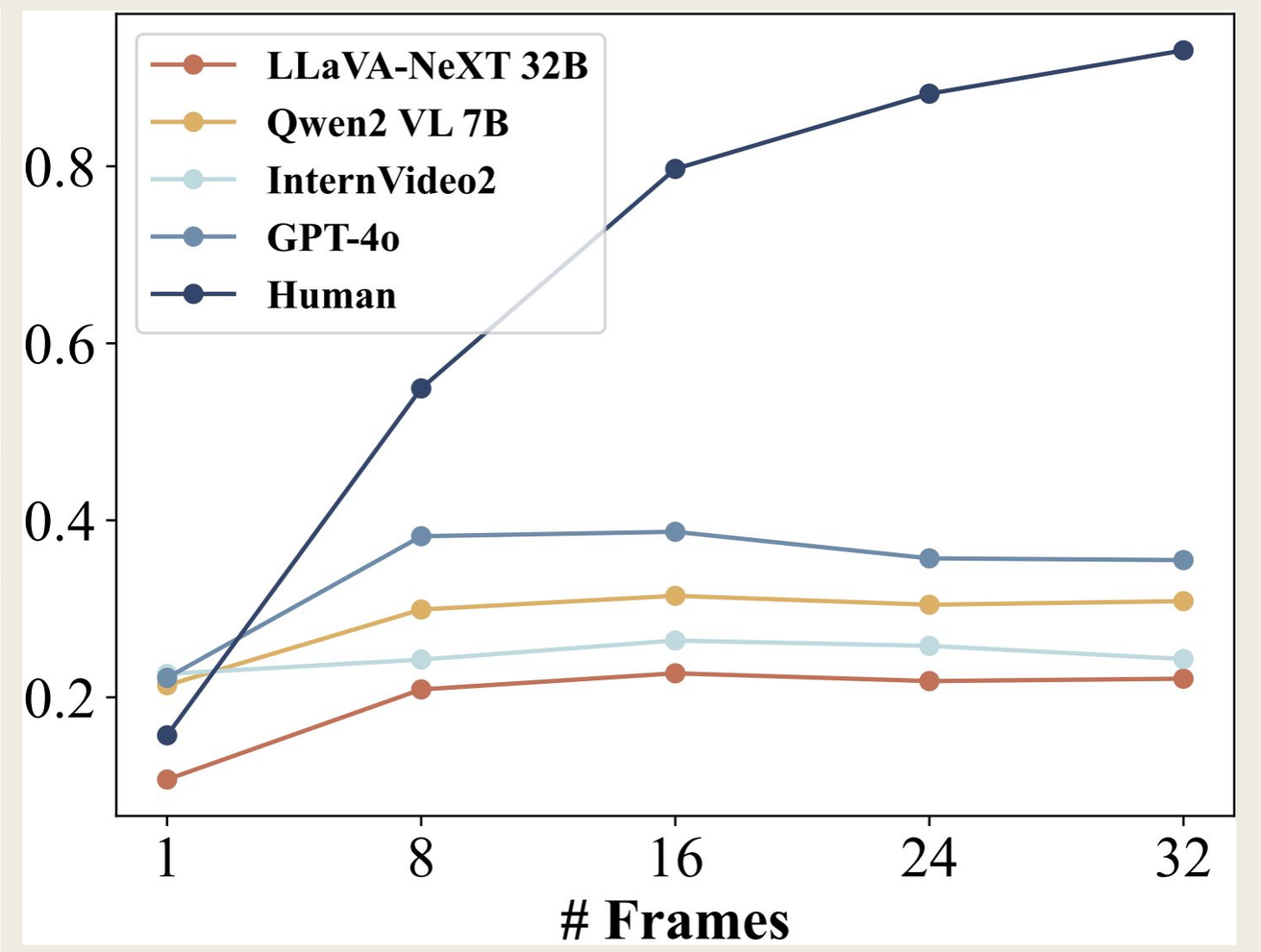
*How much does performance **drop** when we replace the most informative frame with a random one?*

A **lower ρ** value on TOMATO indicates a more even distribution of informativeness across the multiple frames, and the question cannot be more accurately answered by a handpicked 1 frame compared to a random 1 frame.

	VITATECS			MVBench			TempCompass			ReXTime			TOMATO (200)		
# Frames	1[R]	1[H]	ρ ↓	1[R]	1[H]	ρ ↓	1[R]	1[H]	ρ ↓	1[R]	1[H]	ρ ↓	1[R]	1[H]	ρ ↓
Average	70.7	86.0	21.6	47.1	58.8	25.0	51.3	64.0	24.9	62.8	86.1	37.2	21.8	22.8	**4.6**

Main Results

Model	Score
Baselines	
Human (Videos)	95.2
Human (Frames)	79.7
Random (GPT-4o)	23.1
Proprietary Multimodal Foundation Models (MFMs)	
GPT-4o	**37.7**
Gemini 1.5 pro ‡	36.1
Claude 3 Opus	28.2
Claude 3.5 Sonnet	27.8
Open-Source Multimodal Foundation Models (MFMs)	
Qwen2-VL-72B	**37.9**
Qwen2-VL-7B	31.5
Video-CCAM-v1.1 14B	30.7
InternVL 2 40B	29.0
LLaVA-Video-72B	28.2
Video-CCAM-v1.1 9B	27.0
InternVideo 2 8B	26.4

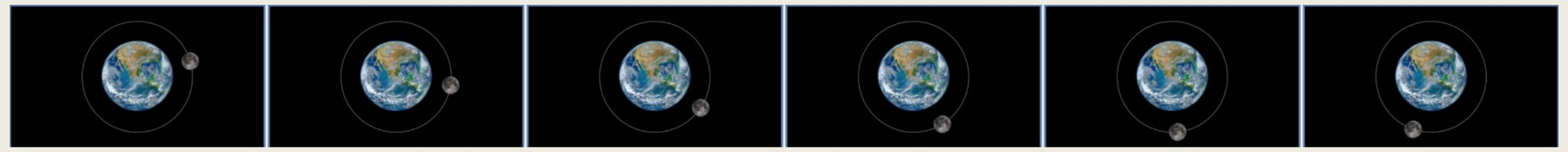


1. **Significant gap** between human and SOTA model performance.

2. **Narrowing gap** between proprietary and open-source models.

Analysis

1. SOTA Models Cannot Interpret Frames as a Continuous Sequence



GPT-4o: … the sequential positions of the moon as it moves from the **right** side of the Earth, **downwards**, to the **left** side, and then **upwards**…
Therefore, the most possible answer based on the frames is:
`A`: Counter-clockwise throughout.

✓ correct description
✗ wrong conclusion

2. SOTA Models are Over-Reliant on Common Sense

GPT-4o: … the object (a cup with liquid) is **initially moving downward and then upward**…the object seems to be **accelerating** as it moves downward due to gravity and **then decelerating** as it is caught or swung back up…
Therefore, the most plausible answer based on the frames is:
`Option D`: Accelerating.

✗ wrong description
✗ wrong conclusion