Lost in Time: A New Temporal Benchmark for VideoLLMs

VIS LAB · QUVA Deep Vision Lab · University of Amsterdam · USC Universidade de Santiago de Compostela · CiTIUS Singular Research Center on Intelligent technologies · Fundamental AI Lab · UTN · BMVC 2025

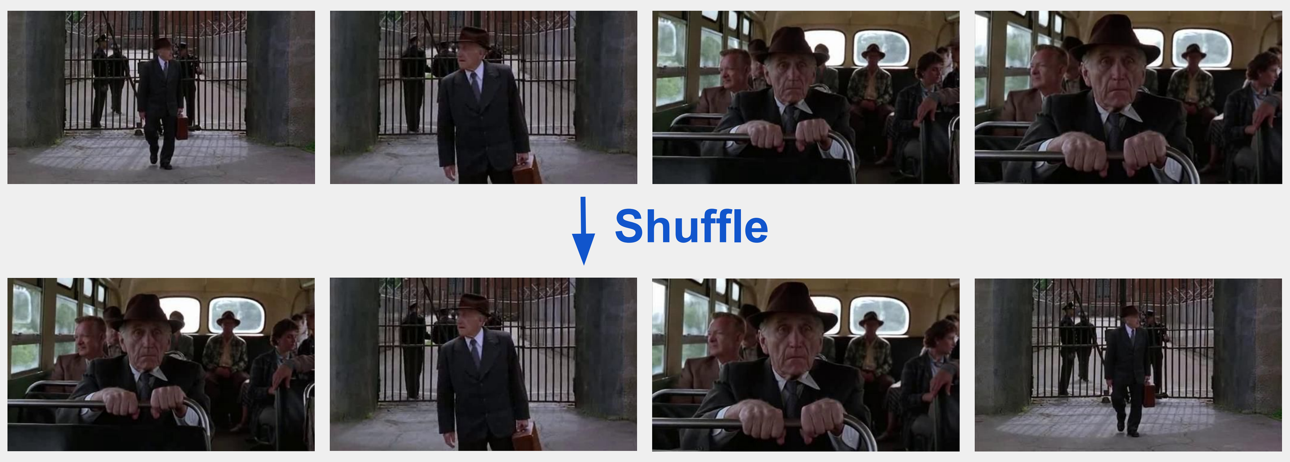
Daniel Cores*, Michael Dorkenwald*, Manuel Mucientes, Cees G. M. Snoek, Yuki M. Asano

Motivation

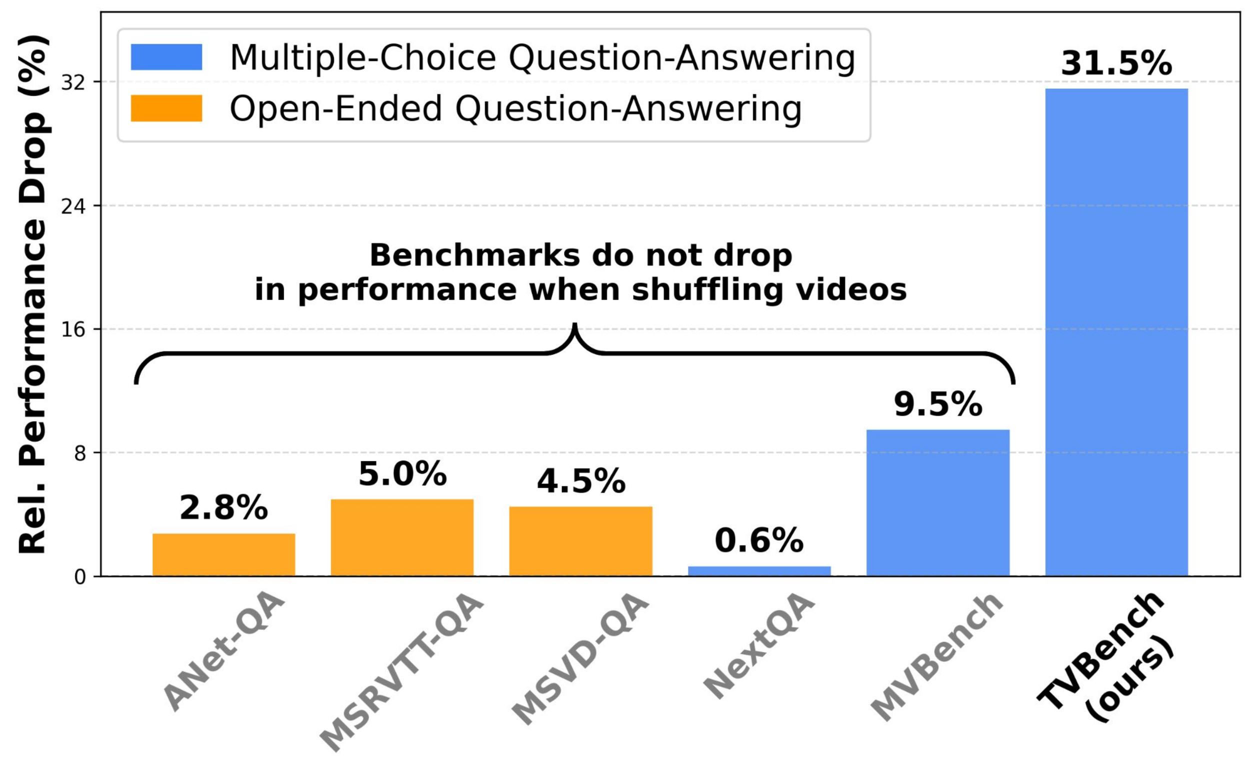
The performance of a SOTA model on commonly used benchmarks remains nearly unchanged when shuffling the input videos.

MVBench

Can you choose the option that matches how the scenes change in the video?



- ✗ From work to the gym.
- ✗ From school to the park.
- ✓ From prison to the bus.
- ✗ From kitchen to the living room.



TVBench

A new benchmark specifically created to evaluate temporal understanding in video QA.

Task	#Videos	Average Frame Length
Action Count	536	628
Object Count	148	127
Action Sequence	437	409
Object Shuffle	225	710
Scene Transition	185	599
Action Localization	160	423
Action Antonym	320	171
Unexpected Action	82	726
Egocentric Sequence	200	562
Moving Direction	232	127

Problems in Video Benchmarks

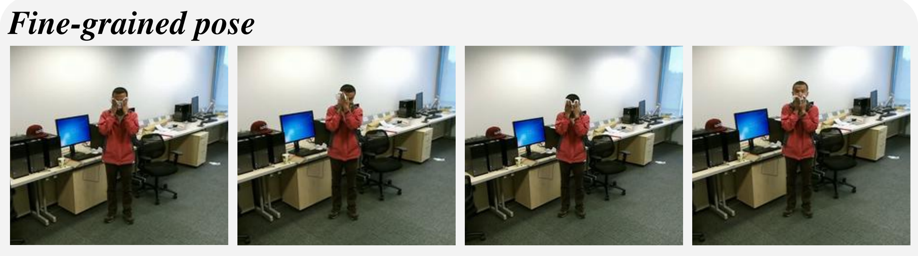
Time Often Does Not Matter

Fine-grained action recognition

What is the action performed by the person in the video?

- ✗ Floating
- ✗ Bubbling
- ✗ Washing
- ✓ Bathing

Fine-grained pose



What is the action performed by the person in the video?

- ✗ Shake head
- ✗ Salute
- ✗ Nod head
- ✓ Wipe face

Answer is obvious from any frame

Vision Often Does Not Matter

What action prevents the ring from getting trapped in the rope?

- ✓ Man avoids letting ring get caught in rope.
- ✗ Man keeps rotating the ring constantly.
- ✗ Man ensures knotting at end of rope.
- ✗ Man throws the ring high in air.

Answer is paraphrased question

Automatic Open-ended evaluation

Q: Who plays a keyboard?
A: Man.

Prediction: ✗ Elton. → GPT-3.5 evaluation → Correct? Yes. Score: 4

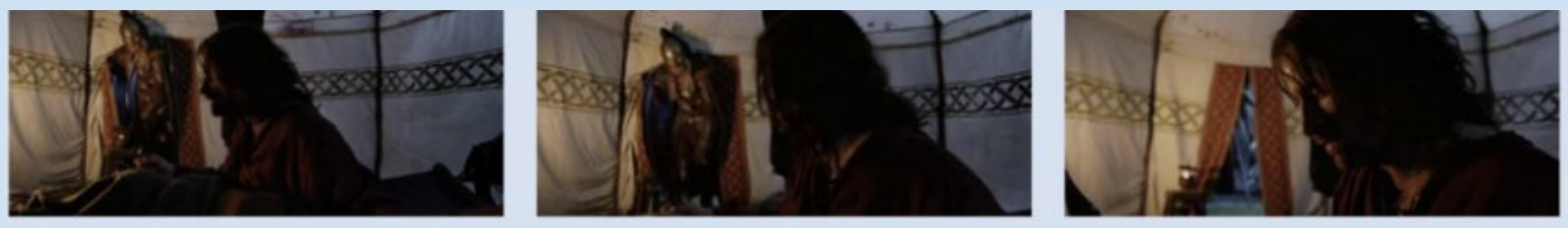
Q: Who rides his motorcycle through the city streets?
A: Man.

Prediction: ✗ Ewan. → GPT-3.5 evaluation → Correct? No. Score: 0

Solutions in TVBench

Temporally hard MCQA candidates

Which direction does the yellow sphere move in the video?



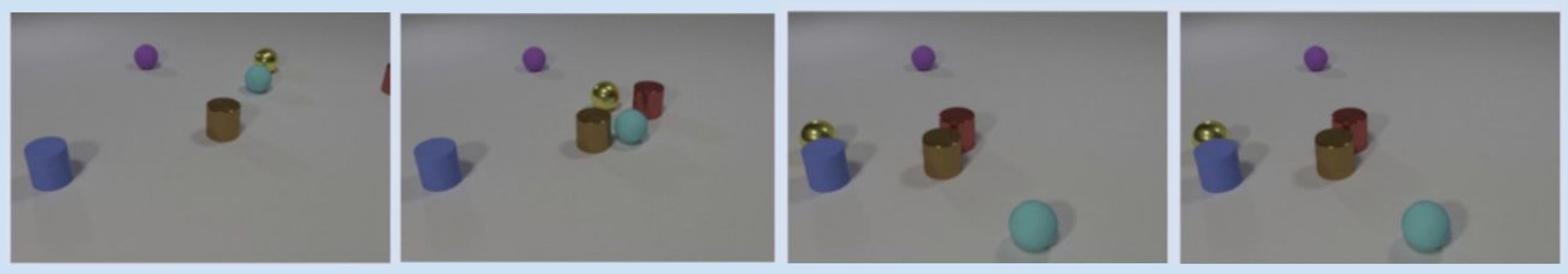
- ✗ From outside the tent to inside the tent.
- ✓ From inside the tent to outside the tent.

What is the action being performed in the video?

- ✗ Take off a hat.
- ✓ Put on a hat.

Carefully designed MCQA templates

Which direction does the *yellow sphere* move in the video?



- ✗ Up and to the right.
- ✗ Up and to the left.
- ✓ Down and to the left.
- ✗ From the bus to prison.

Results

	Model	Input	GPT-3.5 Acc.	GPT-3.5 Score	Llama3-70B Acc.	Llama3-70B Score	ΔAcc.
MSVD	Llama3 70B	text	29.1	2.6	51.9	2.8	+22.8
	GPT-4o	text	29.0	2.5	44.2	2.4	+15.2
	GPT-4o	image	60.6	3.6	75.3	3.8	+14.7
	Tarsier-34B	shuffle	76.7	4.0	78.3	4.1	+1.6
	Tarsier-34B	video	80.3	4.2	78.3	4.1	-2.0
MSRVTT	Lama3 70B	text	23.9	2.4	47.8	2.6	+23.9
	GPT-4o	text	23.3	2.3	42.4	2.3	+19.1
	GPT-4o	image	34.2	2.7	50.8	2.7	+16.6
	Tarsier-34B	shuffle	63.1	3.5	62.7	3.4	-0.4
	Tarsier-34B	video	66.4	3.7	63.0	3.4	-3.4
ActivityNet	Lama3 70B	text	25.3	2.6	32.6	1.9	+7.3
	GPT-4o	text	27.1	2.5	33.9	1.8	+6.8
	GPT-4o	image	46.4	3.2	56.2	2.9	+9.8
	Tarsier-34B	shuffle	59.9	3.6	60.8	3.4	+0.9
	Tarsier-34B	video	61.6	3.7	61.3	3.4	-0.3

Standard deviation between GPT-3.5 and Llama3 score differences: ± 9.3

- Video information is **NOT** crucial to solve current benchmarks.
- Open-ended evaluation is **biased** by the evaluation model.
- Video information **IS** crucial to solve TVBench.

Model	Input	MVBench Avg.	TVBench Avg.
Random	–	27.3	33.3
GPT-3.5 Turbo	text-only	$35.0_{\uparrow 7.7}$	$33.1_{\downarrow 0.2}$
Llama 3 70B	text-only	$38.1_{\uparrow 10.8}$	$34.7_{\uparrow 1.4}$
GPT-4o	text-only	$34.8_{\uparrow 7.5}$	$33.8_{\uparrow 0.5}$
Gemini 1.5 Pro	text-only	$38.2_{\uparrow 10.9}$	$33.6_{\uparrow 0.3}$
Tarsier-34B	text-only	$35.7_{\uparrow 8.4}$	$34.4_{\uparrow 1.1}$
Idefics3	image	$44.2_{\uparrow 16.9}$	$34.5_{\uparrow 1.2}$
GPT-4o	image	$47.8_{\uparrow 20.5}$	$35.8_{\uparrow 2.5}$
Gemini 1.5 Pro	image	$48.5_{\uparrow 21.2}$	$36.3_{\uparrow 3.0}$
Tarsier-34B	image	$45.1_{\uparrow 17.8}$	$35.0_{\uparrow 1.7}$
VideoChat2	video shuffle	$49.8_{\uparrow 22.5}$	$34.7_{\uparrow 1.4}$
PLLaVA-34B	video shuffle	$56.7_{\uparrow 29.4}$	$37.2_{\uparrow 3.9}$
Gemini 1.5 Pro	video shuffle	$56.8_{\uparrow 29.5}$	$36.1_{\uparrow 2.8}$
Tarsier-34B	video shuffle	$61.2_{\uparrow 33.9}$	$38.0_{\uparrow 4.7}$
GPT-4o	video	$49.1_{\uparrow 21.8}$	$39.9_{\uparrow 6.6}$
Gemini 1.5 Pro	video	$60.5_{\uparrow 33.2}$	$47.6_{\uparrow 14.3}$
Qwen2-VL 72B	video	$73.6_{\uparrow 46.3}$	$52.7_{\uparrow 19.4}$
PLLaVA-34B	video	$58.1_{\uparrow 30.8}$	$42.3_{\uparrow 9.0}$
Tarsier-34B	video	$67.6_{\uparrow 40.3}$	$55.5_{\uparrow 22.2}$
Aria	video	$69.7_{\uparrow 42.4}$	$51.0_{\uparrow 17.7}$
Human Baseline	video	-	$94.8_{\uparrow 61.5}$