

Not Another Text Benchmark

Putting the "Visual" Back in Visual Question Answering for Large Video Models

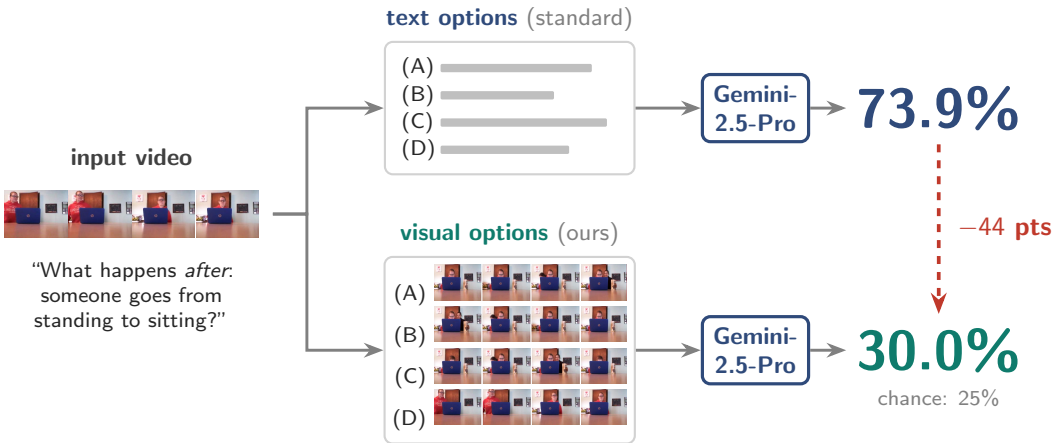
Rwiddhi Chakraborty[1] Yinong Oliver Wang[2] Cheng Zhang[3] Fan Bai[4] Zhuoran Yu[5]
Michael Kampffmeyer[6] Yong Jae Lee[5] Fernando De la Torre[2] Robert Jenssen[6]

[1]University of Copenhagen [2]Carnegie Mellon University [3]Texas A&M University [4]Johns Hopkins University
[5]University of Wisconsin–Madison [6]UiT The Arctic University of Norway

NeurIPS 2026 · Evaluations & Datasets Track

rwchakra.github.io/not-another-text-benchmark

The problem: video benchmarks answer in text



Same questions, same model. Only the answer options change:
frames from the video instead of **text descriptions**.

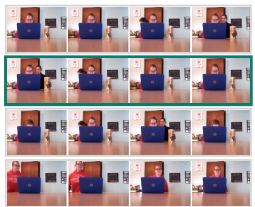
NATB-V: three tasks where the answer is visual

1 Temporal frame retrieval

Charades · 321 videos · 1,106 questions

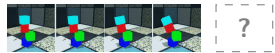


What happens *after* the action?

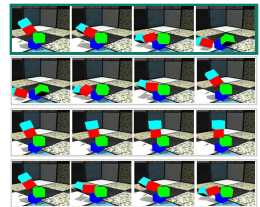


2 Video future prediction

ShapeStacks simulator · 300 videos



What happens *next*?



3 Causal memory distortion

Charades · 20 videos × 6 variants · yes / no



the removed contact frame

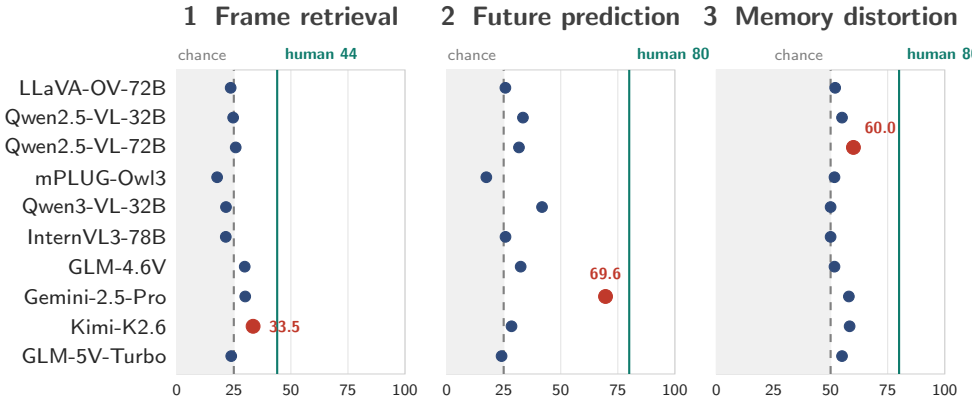


Did this frame appear in the video?

Humans falsely "remember" it when later frames *imply* it (Strickland & Keil, 2011). Do video models?

Direct visual querying instead of text · 641 videos in total.

Results: most models sit near chance

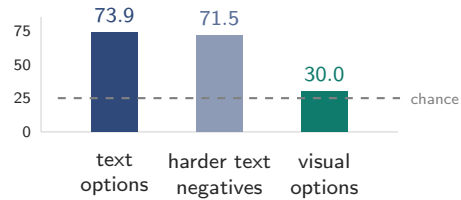


Shaded: at or below chance. **Red**: best model per task.

Is it the visual options? Yes, the modality is the gap

1 Text options inflate accuracy

Frame retrieval, Gemini-2.5-Pro

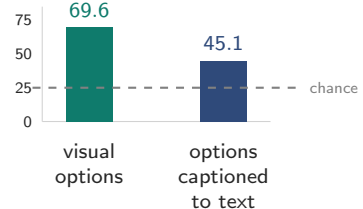


harder negatives VideoCon distractors: 73.9 → 71.5. The gap stays.

reworded prompts text−visual gap −39 to −46 pts (Gemini, Kimi-K2.6)

2 Captioning does not help either

Future prediction, Gemini-2.5-Pro



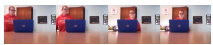
captioned GPT-5.4 describes each option in text

lost fine-grained motion that tells the options apart: −24.5 pts

Text can make video QA look **easier** (given text options) or **harder** (text made from frames). Visual options test the frames directly.

Model failures

1 Gemini-2.5-Pro: option grounding

sees 

reasons describes the events correctly ✓

options

A 

B 

C 

D 

its reasoning: B ✓

its answer: C ✗

Right reasoning, mapped to the wrong frames.

2 Qwen2.5-VL-72B: query grounding

sees 

reasons lists the events, ∼66% correct ✓

options

A 

B 

C 

D 

never compares the options

its answer: C ✗

Understands the video, ignores the visual query.

Takeaway: frontier video models reason well *about* a video in text, but struggle to ground that reasoning in **frames**. NATB-V adds this visual check next to text benchmarks.

Project page: rwchakra.github.io/not-another-text-benchmark

Thank you!

Come talk to us at the poster.

when **Fri, Dec 11, 2026**
11:30 AM – 1:30 PM CET

where **Poster Hall, Paris**

project page

rwchakra.github.io/not-another-text-benchmark

NEURAL INFORMATION PROCESSING SYSTEMS