NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Towards Video Thinking Test (Video-TT)

A Holistic Benchmark for Advanced Video Reasoning and Understanding

Yuhao Dong 董宇昊

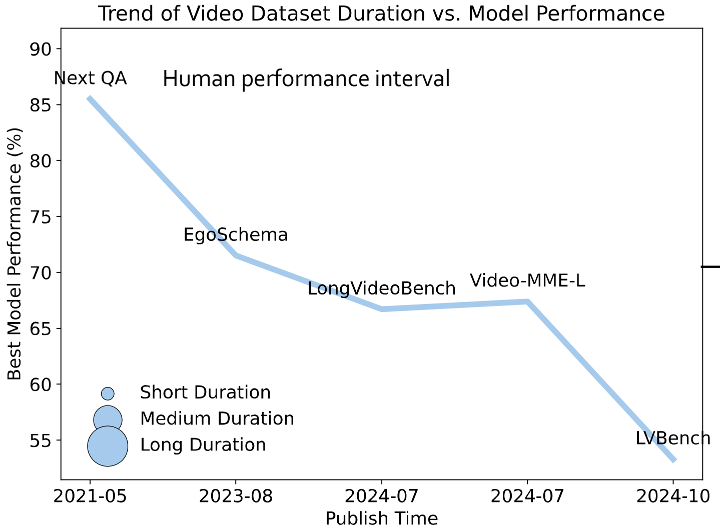
Nanyang Technological University

MMLab@NTU

Video Thinking Test - Motivation

The overall goal of the video understanding benchmark:

To reflect the <u>gap in video understanding</u> between humans and models.



Current development of benchmarks:

- **Short** videos ➡️ **Long** videos,
- **Small** gap ➡️ **Large** gap.

👍 Sounds good!

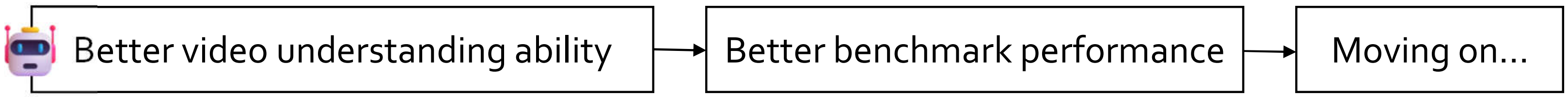
But are we moving at the <u>right pace</u> in building <u>video understanding</u> benchmark?

Video Thinking Test - Motivation

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

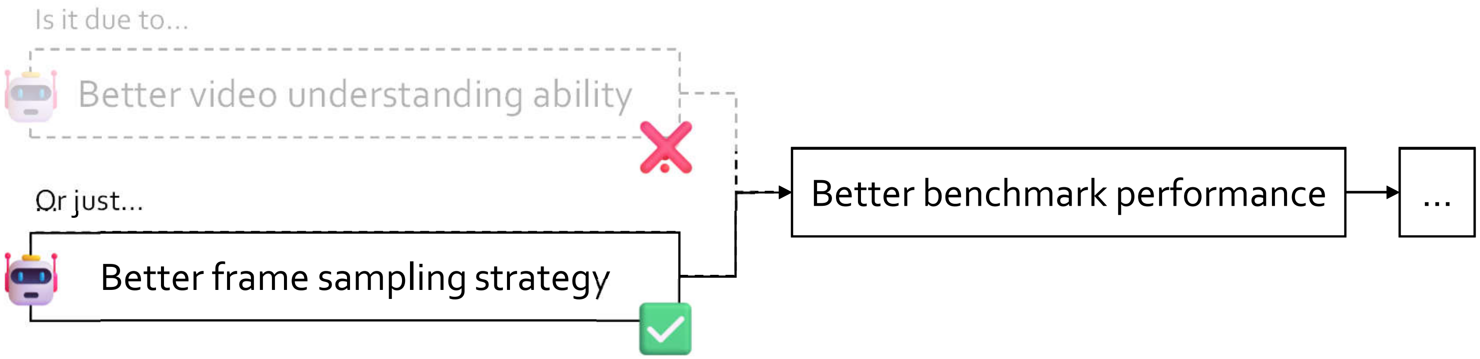
Are we moving at the right pace in building video understanding benchmark?

🌟 Ideal Pace:



Better video understanding ability → Better benchmark performance → Moving on…

😳 Current Pace:



Is it due to…

Better video understanding ability ✗

Or just…

Better frame sampling strategy ✓

→ Better benchmark performance → …

Video Thinking Test - Motivation

Are we moving at the right pace in building video understanding benchmark?

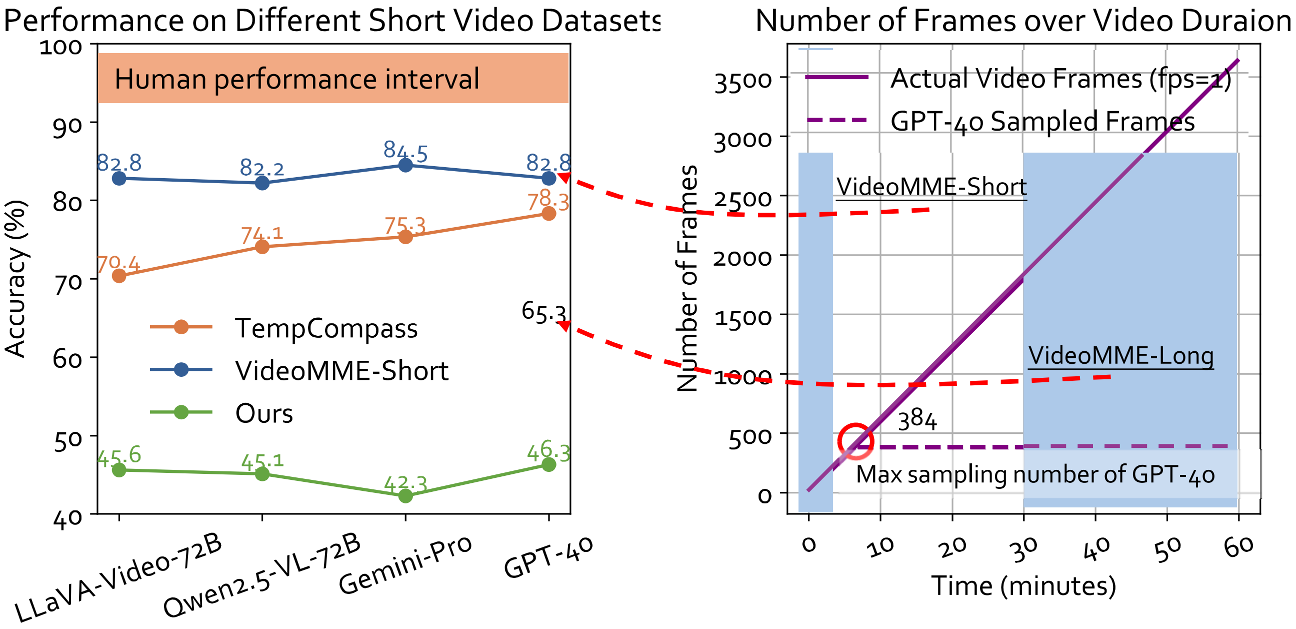
Maybe not ——

Better benchmark performance

is more so due to:

✅ Better frame sampling strategy

❌ Better video understanding ability



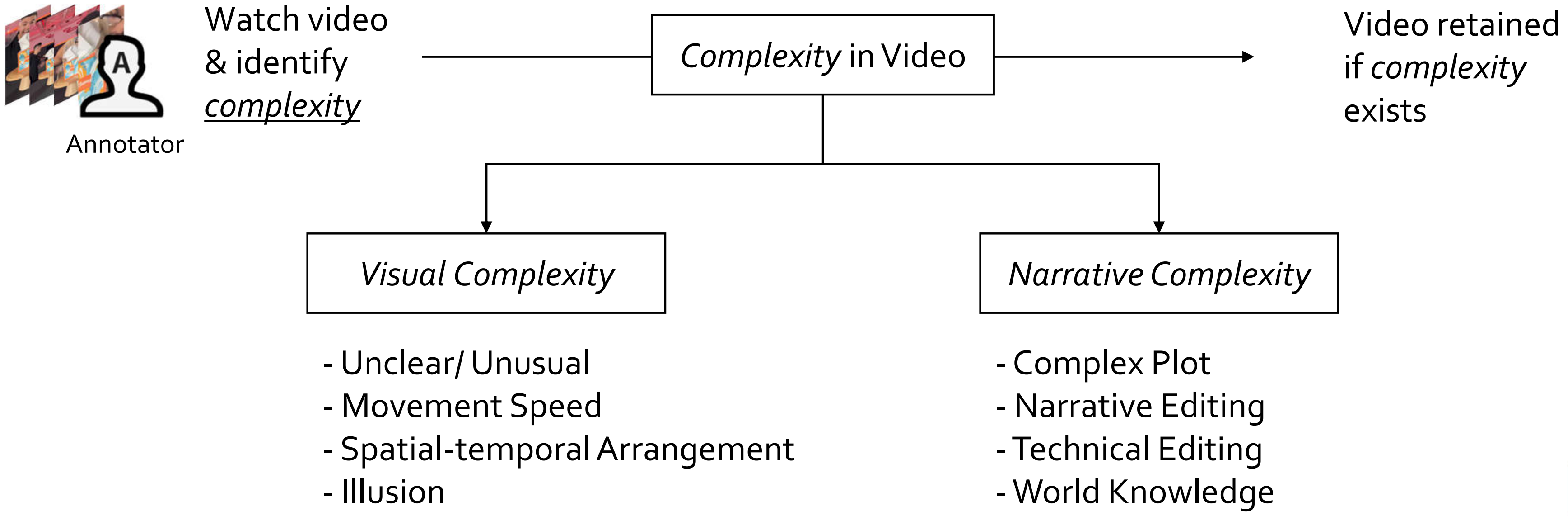
Video Thinking Test - Goal

Reflecting the <u>gap in video understanding</u> ability between humans and models

- **Challenging enough to reveal the human-model performance gap**
  - Ensuring challenge by human-model synergy
  - Annotation with rational and answer
- **Disentangles video understanding from frame sampling**
  - Sampling check

Video Thinking Test - Overview

Ensuring Complexity | Primary Q Annotation | Answer and Rationale Annotation | Sampling Check | Adversarial Q Expansion



Annotator

Watch video & identify *complexity*

Complexity in Video

Video retained if *complexity* exists

Visual Complexity

- Unclear/ Unusual
- Movement Speed
- Spatial-temporal Arrangement
- Illusion

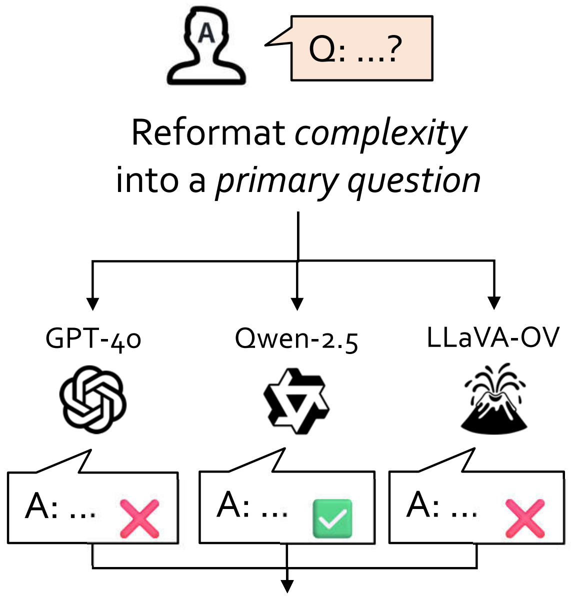
Narrative Complexity

- Complex Plot
- Narrative Editing
- Technical Editing
- World Knowledge

Video Thinking Test - Overview

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE | S-LAB FOR ADVANCED INTELLIGENCE

Ensuring Complexity → **Primary Q Annotation** → Answer and Rationale Annotation → Sampling Check → Adversarial Q Expansion



Reformat *complexity* into a *primary question*

Retained if model response is wrong >1

Example

Narration Complexity — Word Knowledge:
What does the cake look like when it is being eaten?

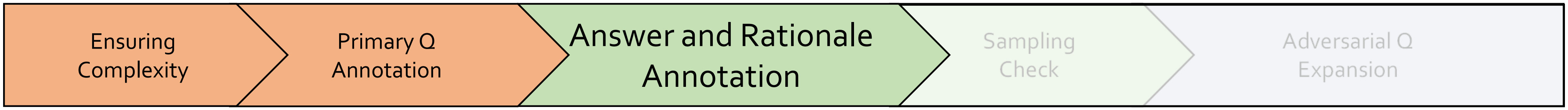
Primary Question:
Is the tissue box real or cake?

Model Response:

- ✗ A real box with tissue paper.
- ✓ It is a cake.
- ✗ It is a real tissue box.

Retained

Video Thinking Test - Overview



Key Conclusion

Human Logical Rationale

Model Wrongness Clarification

Provides a *complete answer with rationale*

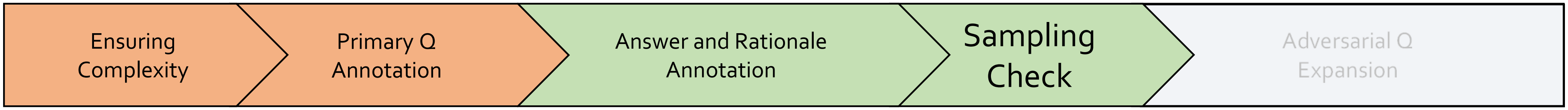
Example

Narration Complexity — Word Knowledge:
What does the cake look like when it is being eaten?

Primary Question:
Is the tissue box real or cake?

Ground Truth Answer & Rationale:
The tissue box is cake.
Because we can see that the tissue box has been bitten by a person, and the shape it reveals is the same as the shape of a cake that has been bitten off.

Video Thinking Test - Overview



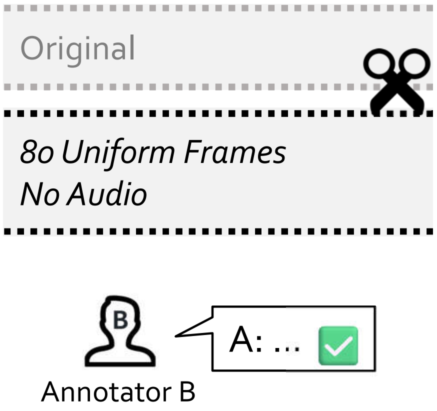
Ensuring Complexity

Primary Q Annotation

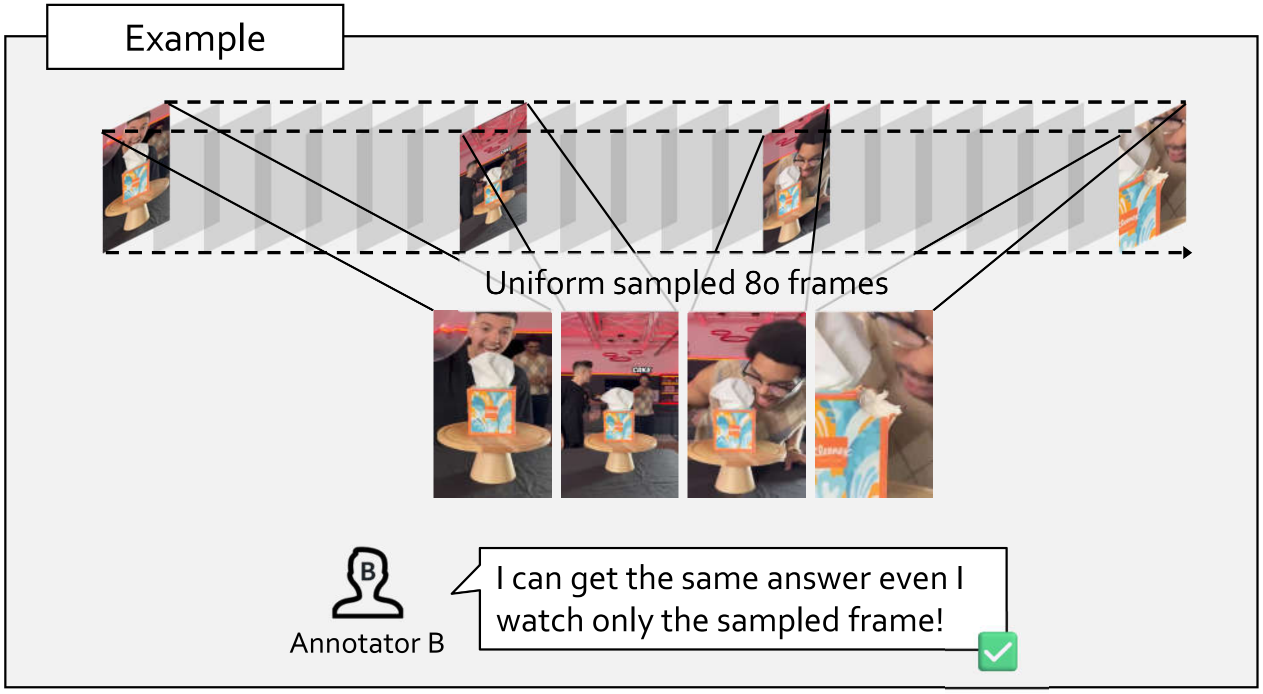
Answer and Rationale Annotation

Sampling Check

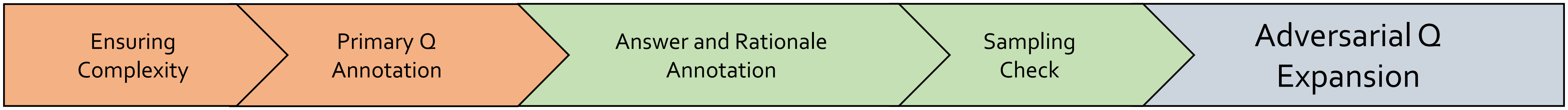
Adversarial Q Expansion

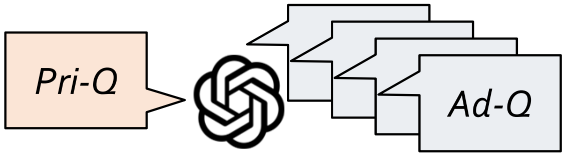


Verify if *answer* can still be reached by humans with less sampling, retained if yes.


Example
Uniform sampled 80 frames
Annotator B
I can get the same answer even I watch only the sampled frame! ✓

Video Thinking Test - Overview


Ensuring Complexity
Primary Q Annotation
Answer and Rationale Annotation
Sampling Check
Adversarial Q Expansion



Primary Question is expanded by GPT-4o into four corresponding *Natural Adversarial Question*,

Then checked and refined by humans.

Example

🤔 **Primary Question:**
Is the tissue box real or cake?

😈 **Rephrased Q:**
What is the tissue box truly?

😈 **Multi-Choice Q:**
What is the tissue box truly?
A. Cake, B. Real box,
C. Plastic, D. Paper.

😈 **Correctly-led Q:**
Is the tissue box a cake?

😈 **Wrongly-led Q:**
Is the tissue box real?

Video Thinking Test – Error Analysis

My daughter's room has this blank wall.
Let's give it a little something special...

Task & Complexity:

Elements Counting → Visual Complexity → *Spatial-Temporal Arrangement — Same Elements in Multiple Frames*

Example

Primary Question:
How many picture frames are showing?

Answer with rationale:
The video displays 10 frames.
As the camera pans from left to right and then returns left, the frames at the end of the video are the same to those at the beginning.

GPT-4o:
The video shows 12 frames. ❌

Video Thinking Test – Error Analysis

Task & Complexity:

Event Localization → Visual Complexity → *Spatial-temporal Arrangement — Event Sequence*

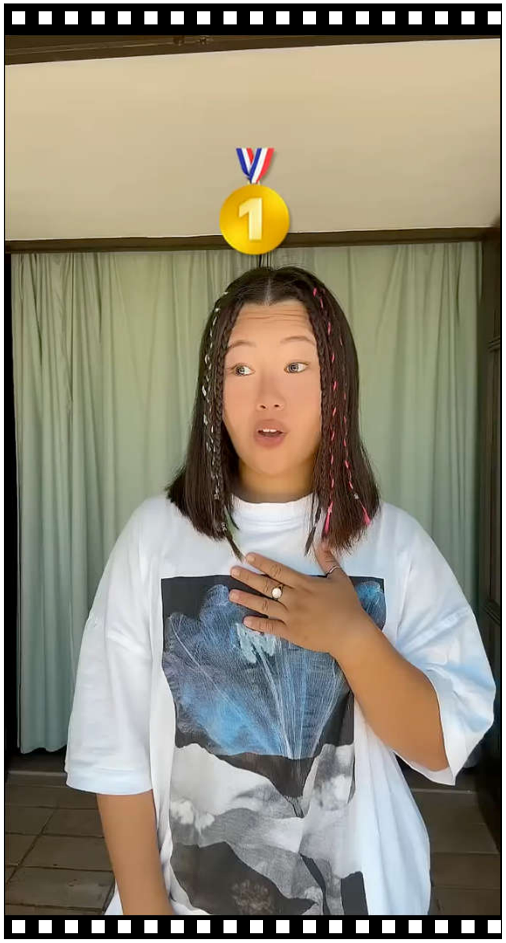
Example

Primary Question:
What are the characteristics of the second person who successfully did a flip in the video?

Answer with rationale:
The second person to attempt a flip is the one wearing a black hoodie. The shirtless man in the first scene tries to flip twice, but fails the first time.

GPT-4o:
The man who is shirtless. ❌

Video Thinking Test – Error Analysis



Task & Complexity:

Character Reaction → Narrative Complexity → *World Knowledge — Psychological Activity*

Example

Primary Question:
What is the reaction of the person who came in second place in the video?

Answer:
She appears to be of disappointment.

GPT-4o:
The person who came in second place appears to be calm and relaxed. ❌

🌟 ***World Knowledge Required***

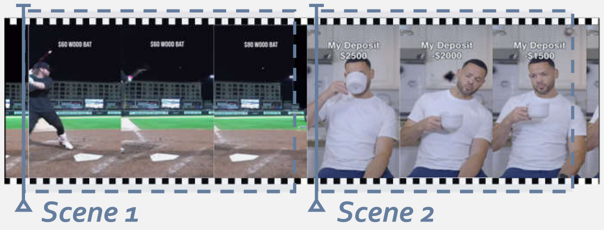
Silver medalists are the least happy, as they narrowly miss gold, while bronze medalists feel relieved to make the podium.

Video Thinking Test – Error Analysis

Task & Complexity:

Plot Attributes → Narrative Complexity → *Complex Plot – In-context Reasoning*

Example



GPT4o: A person is seen at a baseball field, taking swings with various wood bats of different price points. The background includes a large building with advertisements

GPT4o: A person is sitting in a kitchen, holding a white mug. The text on screen reads "My Deposit $2500" initially, and then changes to "-$2000" and finally "-$1500." .The person is seen sipping from the mug and at one point.

👍 *Correct Description to both scene*

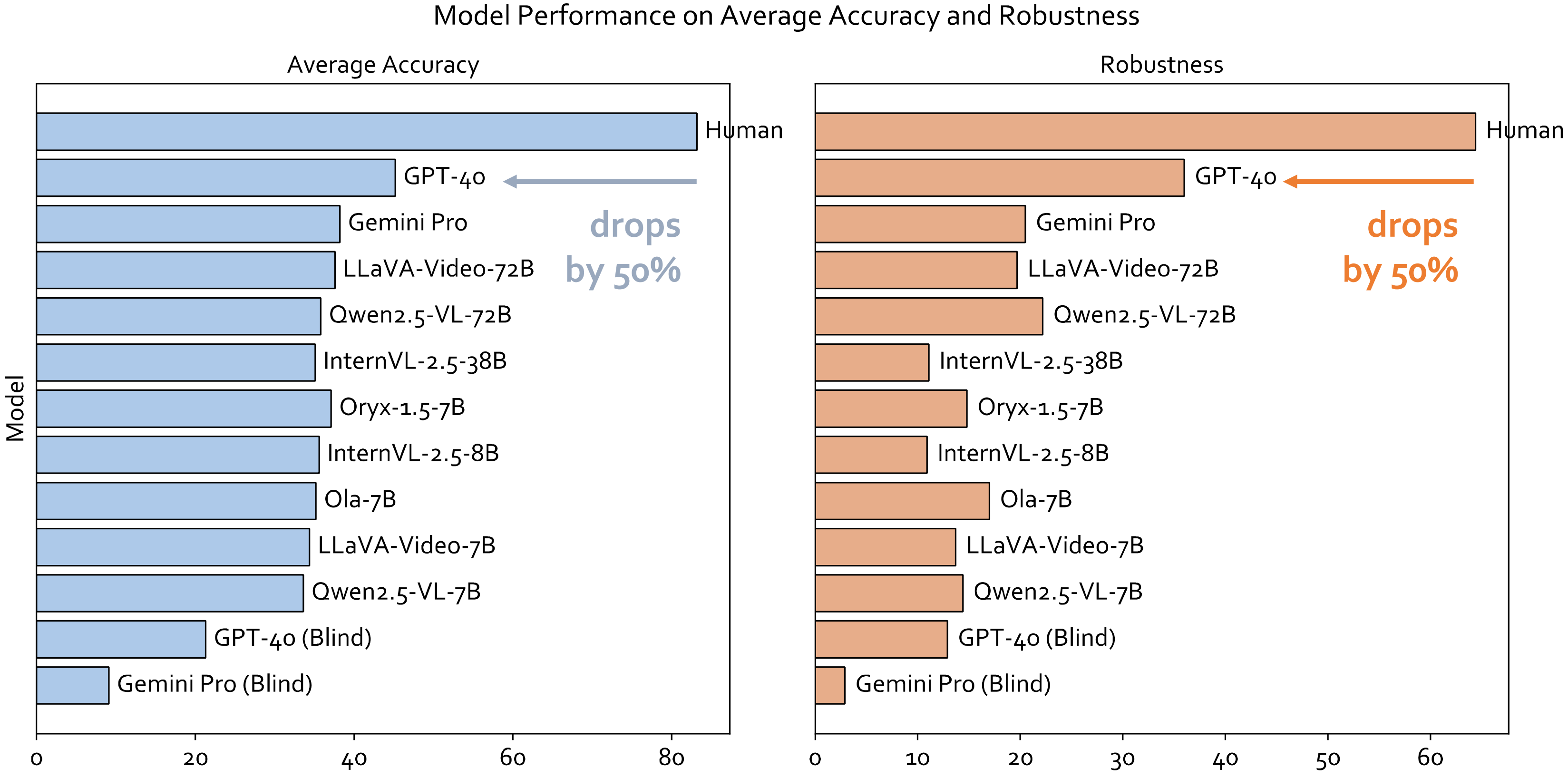
👎 *But fails to link different scenes to create a logical sequence*

Q-4: Combining the different scenes, what is the video trying to imply narratively?

A: The video illustrates the financial impact of damaging a rented house cause by the person playing baseball outside.

GPT-4o: The video shows the amount of money this person spent on baseball games.

Video Thinking Test – Performance



**Please refer to the paper for the definitions and metrics of accuracy and robustness.*

Video Thinking Test – Solutions

Challenge Overview:

- 19 Teams, with 64 final valid submissions.
- Solutions spanning video supervised finetuning, reinforcement learning, multi-agent design and tool-integrated system.
- Achieve new SOTA results (50.7 on multi-choice track), surpassing previous open-source models.

Video Thinking Test – Solutions

Representative Solutions:

Frame Sampling Method:

- Adaptive frame selection at different periods

Reinforcement Learning for Video Reasoning:

- Visual pretext task as training objective: temporal sequence reorder
- Tool-integrated video reasoning: localization, highlighting, etc.

Multi-Agent System:

- Explicit perception/reasoning decomposition

Thank You

Yuhao Dong 董宇昊

Nanyang Technological University

MMLab@NTU