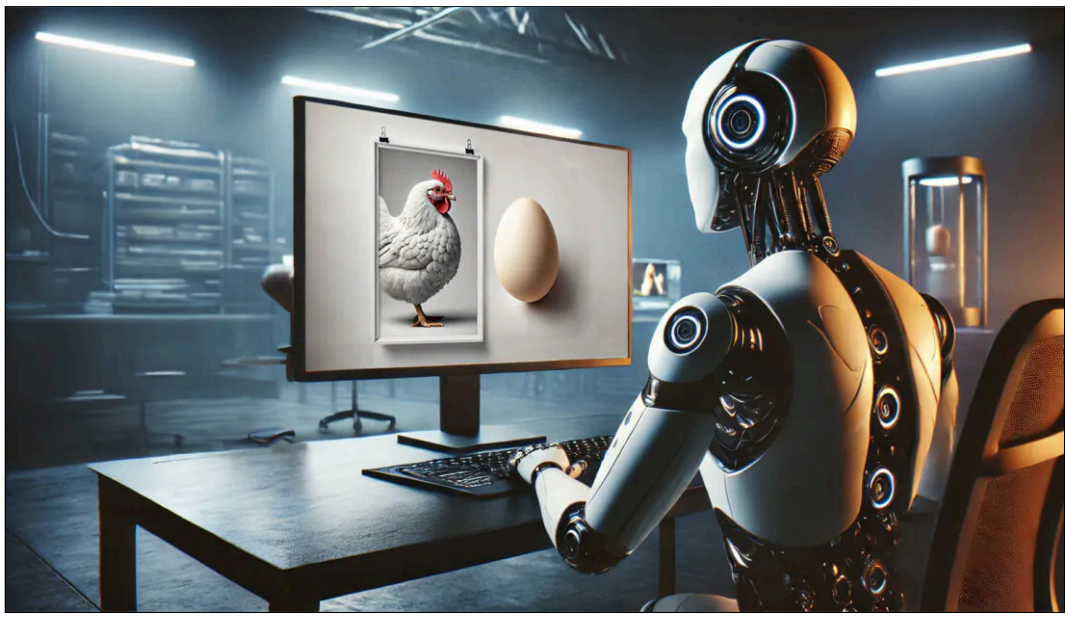
Even State-Of-The-Art Language Models Struggle to Understand Temporal Logic

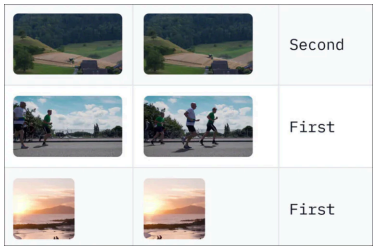
Originally published at https://www.unite.ai/even-state-of-the-art-language-models-struggle-to-understand-temporal-logic/



Predicting future states is a critical mission in computer vision research – not least in robotics, where real-world situations must be considered. Machine learning systems entrusted with mission-critical tasks therefore need adequate understanding of the physical world.

However, in some cases, an apparently impressive knowledge of temporal reality could be deceptive: a new paper from the United Arab Emirates has found that state-of-the-art Multimodal Large Language Models (MLLMs), including sector leaders GPT-4o and Google Gemini, fall short when it comes to interpreting how time is represented in images.

Example sequential pairs (see image below), which would be unchallenging for humans even when put in the wrong order, can fox advanced MLLMs when presented in unexpected contexts or configurations (such as second-image-first, concatenated into single images, sequential multiple images which may or may not represent the correct temporal order, and so on.).



Samples from one of the datasets compiled for the new study, which show sequential events in the form of 'before and after' images. The researchers have made this data available at https://huggingface.co/datasets/fazliimam/temporal-vqa/viewer

The researchers tasked the models with basic temporal reasoning challenges, such as determining event order or estimating time gaps, and found that the seven MLLMs tested performed notably below human accuracy:

> *'Overall, the [results] reveal that all current MLLMs, including GPT-4o – the most advanced model in our evaluation – struggle with the proposed benchmark. Despite GPT-4o's superior performance relative to other models, it fails to consistently demonstrate accurate temporal reasoning across different settings.*
>
> *'The consistent accuracy scores are notably low for all models, indicating significant limitations in their ability to comprehend and interpret temporal sequences from visual inputs. These deficiencies are evident even when models are provided with multiimage inputs or optimized prompts, suggesting that current architectures and training methodologies are insufficient for robust temporal order understanding.'*

Machine learning systems are designed to optimize to the most accurate, but also the most efficient and people-pleasing results*. Since they do not reveal their reasoning explicitly, it can be difficult to tell when they're cheating, or using 'shortcuts'.

In such a case, the MLLM may arrive at the *right answer* by the *wrong method*. The fact that such an answer can be correct may inspire false confidence in the model, which could produce incorrect results by the same method in later tasks presented to it.

Worse yet, this misdirection can become even more deeply embedded in the development chain if humans are impressed by it, and give positive feedback in trials and annotation sessions which may contribute to the direction that the data and/or the model might take.

In this case, the suggestion is that MLLMs are 'faking' a true understanding of chronology and temporal phenomena, by observing and anchoring on secondary indicators (such as time-stamps, for instance, in video data, order of images in a layout, or even – potentially – sequentially-numbered file-names).

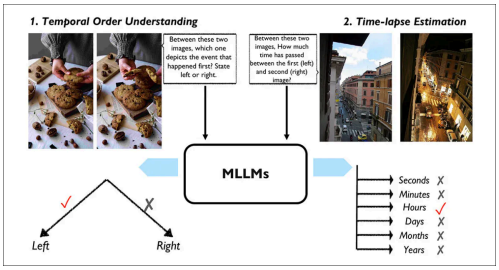
It further indicates that MLLMs currently fail to satisfy any real definition of having generalized a concept of temporal phenomena – at least, to the extent that humans can.

The new paper is titled *Can Multimodal MLLMs do Visual Temporal Understanding and Reasoning? The answer is No!*, and comes from three researchers at the Mohamed bin Zayed University of Artificial Intelligence and Alibaba International Digital Commerce.

Data and Tests

The authors note that prior benchmarks and studies, such as MMMU and TemporalBench, concentrate on single-image inputs or else formulate questions for the MLLMs that may be rather too easy to answer, and may not uncover a tendency towards shortcut behavior.

Therefore the authors offer two updated approaches: *Temporal Order Understanding* (TOU) and *Time-lapse Estimation* (TLE). The TOU approach tests the models on their ability to determine the correct sequence of events from pairs of video frames; the TLE method evaluates the MLLM's ability to estimate the time difference between two images, ranging from seconds to years.

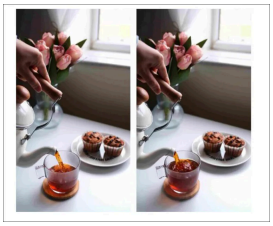


From the paper, the two main tasks of the TemporalVQA benchmark: in Temporal Order Understanding, the model decides which of two images shows an event that occurred first; in Time-lapse Estimation, the model estimates how much time has passed between two images, selecting from options including seconds, minutes, days, or years. These tasks aim to test how well the MLLMs can reason about the timing and sequence of visual events. Source: https://arxiv.org/pdf/2501.10674

The researchers curated 360 image pairs for the TOU benchmark, using open source videos from Pixabay and Pexels, so that it would be possible to make the dataset available via a GUI.

The videos covered a range of subjects, from people in everyday activities to non-human content such as animals and plants. From these, pairs of frames were selected to depict a sequence of events with sufficient variation to make the starting frame 'obvious'.

Human selection was used to ensure that the frames could be definitively ordered. For example, one of the curated pairs shows a partially-filled teacup in one frame, and the same cup fully filled with tea in the next, making the sequence logic easy to identify.



The temporal logic of these two pictures cannot be escaped, since the tea cannot possibly be sucked back up the spout.

In this way, 360 image pairs were obtained.

For the TLE approach, copyright-free images were chosen from Google and Flickr, as well as select frames from copyright-free videos on YouTube. The subject-matter of these videos featured scenes or objects whose change interval ranged from seconds to days to seasons – for example, ripening fruit, or the change of seasons in landscapes.

Thus 125 image pairs were curated for the TLE method.

Not all of the MLLMs tested were able to process multiple images; therefore tests differed to accommodate each model's capabilities.

Multiple versions of the curated datasets were generated, in which some of the pairs were concatenated vertically, and others horizontally. Further variations swapped the true and correct temporal sequence of the pairs.

Two prompt-types were developed. The first followed this template:

> *Did the event in the* ***(left / top / first)*** *image happen before the event in the* ***(right / bottom / second)*** *image? State true or false with reasoning.*

The second followed this schema:

> *Between these two images, which one depicts the event that happened first? State* ***(left or right / top or bottom / first or second)*** *with reasoning.*

For TLE, questions were multiple-choice, asking the models to evaluate the time-lapse between the two presented images, with *seconds*, *hours*, *minutes*, *days*, *months* and *years* available as the time-units. In this configuration, the most recent image was presented on the right.

The prompt used here was:

> *In the given image, estimate the time that has passed between the first image (left) and the second image (right).*
>
> *Choose one of the following options:*
>
> 1. ***Less than 15 seconds***
>
> ***B. Between 2 minutes to 15 minutes***
>
> ***C. Between 1 hour to 12 hours***
>
> ***D. Between 2 days to 30 days***
>
> ***E. Between 4 months to 12 months***
>
> ***F. More than 3 years***

The MLLMs tested were ChatGPT-4o; Gemini1.5-Pro; LlaVa-NeXT; InternVL; Qwen-VL; Llama-3-vision; and LLaVA-CoT.

Temporal Order Understanding: Results

	GPT-4o		Gemini1.5-Pro		LLaVA-NeXT		InternVL		Qwen-VL		Llama3-Vision		LLaVA-CoT	
	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2	P1	P2
Vertically Joined														
Correct answer on the top (1st order)	93.9%	94.7%	30.8%	81.4%	25.3%	100%	71.4%	85.3%	6.1%	44.4%	6.4%	5.25%	84.2%	91.4%
Correct answer on the bottom (2nd order)	32.2%	40.6%	88.3%	41.9%	71.9%	0.0%	27.2%	17.8%	96.9%	66.4%	93.6%	46.1%	16.7%	13.3%
Average Accuracy	63.1%	67.6%	59.6%	61.7%	48.6%	50.0%	49.3 %	51.5%	51.5%	55.4%	50.0%	49.3%	50.4%	52.4%
Consistent Accuracy	31.4%	39.2%	23.6%	35.0%	6.4%	0.0%	16.4%	13.6%	5.8%	19.2%	3.3%	7.8%	13.1%	12.8%
Horizontally Joined														
Correct answer on the left (1st order)	95.0%	92.8%	52.5%	70.8%	7.5%	100%	66.4%	79.4%	10.8%	100%	3.3%	33.3%	83.6%	89.7%
Correct answer on the right (2nd order)	29.4%	49.4%	62.8%	44.4%	89.2%	0.0%	39.4%	22.8%	92.8%	0.0%	95.6%	64.2%	18.6%	12.5%
Average Accuracy	62.2%	71.1%	57.6%	57.6%	48.3%	50.0%	52.9%	51.1%	51.8%	50.0%	49.4%	48.8%	51.1%	51.1%
Consistent Accuracy	30.6%	52.8%	21.4%	29.4%	4.2%	0.0%	25.6%	18.6%	8.3%	0.0%	1.9%	4.4%	14.2%	10.3%
Separate Multiple Images														
Correct answer is the first image (1st order)	81.9%	85.6%	-	-	0.3%	100%	72.5%	67.2%	0.6%	50.8%	-	-	35.0%	94.2%
Correct answer is the second image (2nd order)	54.2%	70.8%	-	-	99.4%	0.0%	28.6%	40.8%	99.4%	59.4%	-	-	69.2%	6.7%
Average Accuracy	68.1%	78.2%	-	-	49.9%	50.0%	50.6%	54.0%	50.0%	55.1%	-	-	52.1%	50.4%
Consistent Accuracy	43.6%	65.3%	-	-	0.0%	0.0%	22.2%	26.4%	0.6%	23.9%	-	-	14.7%	4.7%

Results of Temporal Order Understanding across different models and input layouts, showing accuracy and consistency for various setups and prompts.

Regarding the results shown above, the authors found that all tested MLLMs, including GPT-4o (which showed the best overall performance), struggled significantly with the TemporalVQA benchmark – and even GPT-4o failed to consistently exhibit reliable temporal reasoning across different configurations.

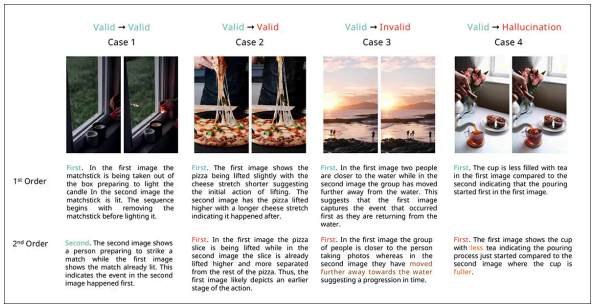
The authors contend that the consistently low accuracy across LLMs highlights significant shortcomings in the models' ability to interpret and reason about temporal sequences from visual data. The researchers note that these challenges persist even with the use of multi-image inputs and optimized prompts, pointing to fundamental limitations in current model architectures and training methods.

The tests showed significant variations in performance across prompting strategies. While GPT-4o improved with optimized prompts (reaching 4% in single-image and 65.3% in multi-image settings), performance remained below acceptable levels.

Models such as LLaVA-NeXT and Qwen-VL were even more sensitive, with performance declining when alternate prompts were used, suggesting that prompt engineering alone cannot overcome the MLLMs' fundamental limitations in regard to temporal reasoning.

Tests also indicated that image layout (i.e., vertical vs. horizontal) significantly impacted model performance. GPT-4o improved its consistency with vertical arrangements, rising from 39.2% to 52.8%; however, other models, including the LLaVA strains, showed strong directional biases, excelling in one orientation but failing in another.

The paper indicates that these inconsistencies suggest reliance on spatial cues, rather than true temporal reasoning, with the MLLMs not genuinely analyzing the sequence of events or understanding the progression over time. Instead, they appear to have relied on patterns or visual features related to the layout of images, such as their position or alignment, in order to make decisions.



Qualitative tests highlights GPT-4o's predictions when faced with different input orders. In the first order, image pairs are presented in their original sequence, while in the second order, the sequence is reversed. Correct classifications are marked in green, pure misclassifications in red, hallucinated reasoning in orange, and illogical or 'invalid' reasoning in brown, revealing the model's inconsistencies across different input configurations.

Comparison tests between single-image and multi-image inputs demonstrated limited overall improvement, with GPT-4o performing slightly better on multi-image input, rising from 31.0% to 43.6% (with P1) and 46.0% to 65.3% (with P2).

Other models, such as InternVL, demonstrated stable but low accuracy, while Qwen-VL saw minor gains. The authors conclude that these results indicate that additional visual context does not substantially enhance temporal reasoning capabilities, since models struggle to integrate temporal information effectively.

Human Study

In a human study, three surveys were conducted to assess how closely the best-performing multimodal MLLM perfgormed against human estimation.

Humans achieved 90.3% accuracy, outperforming GPT-4o's 65.3% by 25%. The dataset proved reliable, with minimal human errors and consistent agreement on correct answers.

Best Model (GPT-4o)	65.3%
Avg Human Evaluation	90.3%

Results from the human user study for the first round of tests.

Time-lapse Estimation: Results

Time Span	Human	GPT-4o	Gemini1.5-Pro	LLaVA-NeXT	Qwen-VL	InternVL	Llama3-Vision	LLaVA-CoT
Seconds	87.7%	89.5%	78.9%	78.9%	63.2%	26.3%	100%	94.7%
Minutes	78.9%	80.0%	90.0%	10.0%	65.0%	75.0%	5.0%	10.0%
Hours	76.5%	40.0%	60.0%	0.0%	30.0%	20.0%	0.0%	30.0%
Days	85.2%	61.1%	44.4%	0.0%	77.8%	11.1%	0.0%	88.9%
Weeks/Months	85.4%	63.2%	47.4%	0.0%	21.1%	5.3%	0.0%	63.2%
Years	86.2%	86.2%	58.6%	24.1%	86.2%	0.0%	24.1%	44.8%
Average Acc.	83.3%	70.0%	63.2%	18.8%	57.2%	22.9%	21.5%	55.3%

Results for TLE: time-lapse estimation evaluates model accuracy in identifying intervals between image pairs, across scales from seconds to years. The task assesses each model's ability to select the correct time scale for the temporal gap.

In these tests, the MLLMs performed only adequately on time-lapse estimation: GPT-4o achieved 70% accuracy, but the other models performed significantly worse (see table above), and performance also varied notably across the various time scales.

The authors comment:

> *'The task of time-lapse estimation tests the ability of MLLMs to infer temporal intervals between image pairs. [All] MLLMs, including top performers like GPT-4o and Gemini1.5-Pro, struggle with this task, achieving only moderate accuracy levels of 60-70%. GPT-4o shows inconsistent performance, with strong performance in Seconds and Years but underperforming in Hours.*
>
> *Similarly, LLaVA-CoT demonstrates exceptional performance in the time spans of Seconds and Days, while showing notably poor performance in the other time intervals.'*

Human Study

In the human study for TLE, average human performance improved on GPT-4o (the best-performing model also in this category) by 12.3%.

The authors note that some of the challenges were particularly demanding, and that in one case all the human participants returned a wrong answer, along with all the AI participants.

The authors conclude that GPT-4o exhibits 'reasonably robust reasoning capabilities, notwithstanding the order of images presented to it.

Conclusion

If MLLMs eventually amass and absorb enough 'shortcut' data to cover even the trickiest challenges of the type presented by the authors in this study, whether or not they can be said to have developed human-style generalization capabilities in this domain could become a moot point.

Neither is it known exactly by what route we obtain our own abilities in temporal reasoning – do we likewise 'cheat' until the sheer quantity of learned experience reveals a pattern that performs as 'instinct' in regards to this kind of test?

* *From the point of view that models are increasingly being optimized with loss functions which human feedback has contributed to, and effectively optimized by human trials and subsequent triage.*

First published Monday, January 27, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai