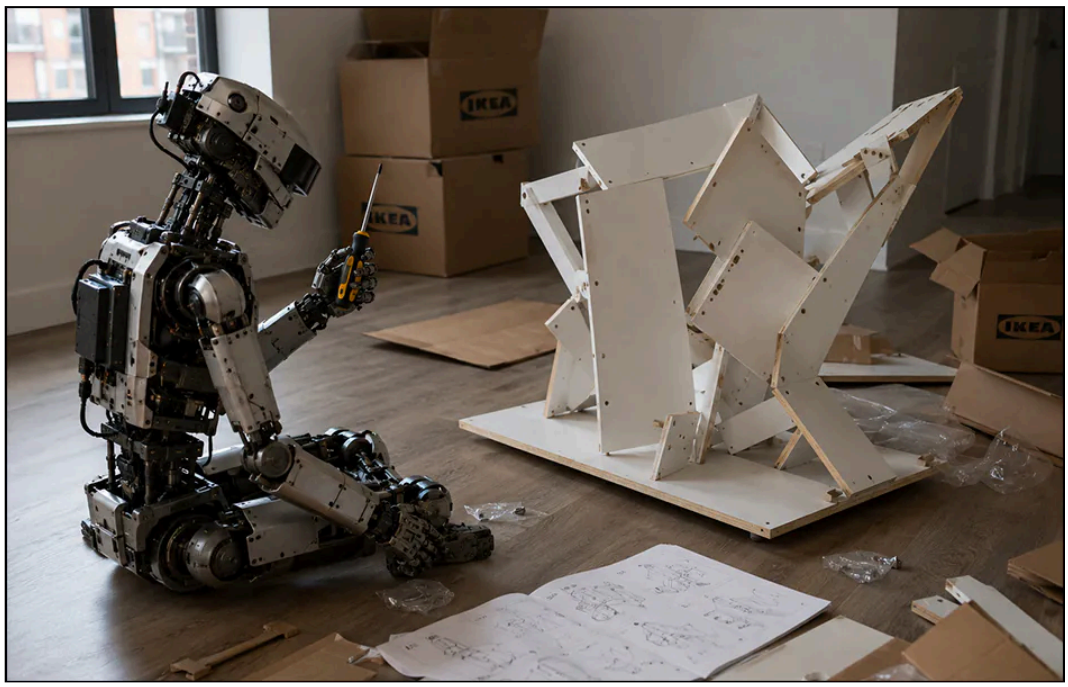
AI Is Significantly Worse Than Humans at Assembling Furniture

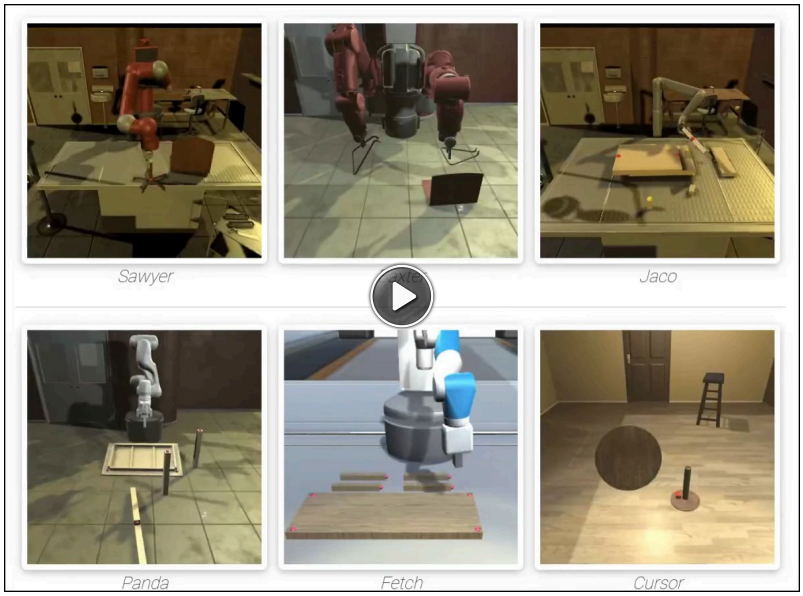
Originally published at https://www.unite.ai/ai-is-significantly-worse-than-humans-at-assembling-furniture/



ChatGPT and Google Gemini still cannot reliably understand IKEA assembly videos, with many other prominent AI systems confusing parts, missing connections, and barely using the video itself to figure out what is happening.

The enduring cultural meme around the difficulty of assembling IKEA-style flat-pack furniture makes the subject an attractive target for computer-vision research — not least because the long sequences of actions, object-tracking, and spatial reasoning involved will tend to push robotic manipulation systems far beyond the simplified shapes and controlled environments to which they are accustomed.

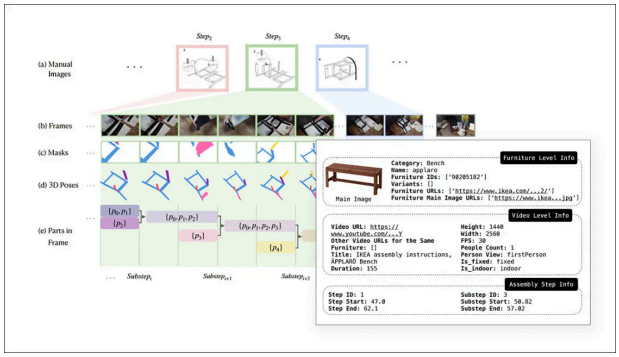
Therefore work on AI-powered robotic assembly routines for flat-packed furniture has become a small but respectable branch in the literature, with outings such as USC's 2019 IKEA Furniture Assembly Environment, among the first benchmark datasets and research contexts specifically aimed at furniture assembly:



CLICK TO PLAY VIDEO ON SITE

Click to play *Examples of robotic assembly practice, from the project site for the 2019 IKEA Furniture Assembly Environment initiative.* Source

In 2024 the Stanford/J.P. Morgan collaboration *IKEA Manuals at Work* was the first to significantly probe AI's capability to undertake this apparently mundane (if often frustrating) procedure, based on a novel dataset of images from instruction manuals, and using instructional videos:



Dataset method and details from the 2024 IKEA Manuals at Work initiative. Source

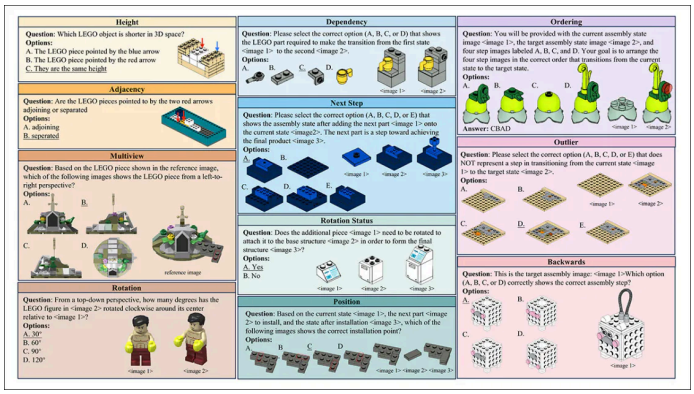
The authors of the 2024 paper – which leveraged DGCNN, CNOS, SAM-6D, MegaPose, MiDaS, SAM2 Hiera-L, Cutie-base, and GPT-4o – concluded that the task yielded *'significant challenges in grounding instructional assembly videos, including extracting part segmentations and poses, constructing high-level assembly plans, and detecting key assembly steps in videos'*.

Wax On, Wax Off

It must be obvious that, while getting AI to automate us out of a task that few cherish would be nice, it's hardly a scientific lodestar, or high in a list of priorities for the Computer Vision research sector.

Rather, the value of the task lies in the fact that what AI systems need to learn *in order to become proficient at this* would ground them for far more serious routines that are equally or even more challenging, in agriculture, industry, the service sector, and diverse other spheres.

In this vein, the LEGO-Puzzles project and dataset examines how well Vision Language Models (VLMs) handle multi-step spatial reasoning across a range of architectures, since assembly tasks depend not only on pairing the correct objects together at the correct moment – a process known as *mating* – but also on following instructions that may be far more abstract than the raw visual scene available to the model at any given step:



Challenging questions from the LEGO-Puzzles project. Source

The latest project to tackle the challenge of furniture assembly exploits a more current and capable crop of AI models, including Google Gemini 2.5/3.1 and OpenAI's GPT-5 – but still fails to obtain a win for AI in the task, with only modest improvements over baseline chance, and performance 'way below human levels'.

The authors state:

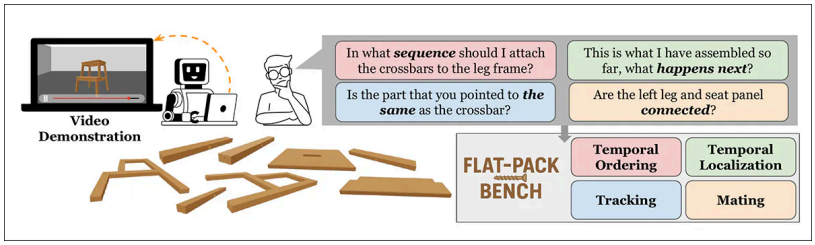
> *'Our experiments reveal that state-of-the-art LVLMs struggle significantly with fine-grained spatio-temporal reasoning, highlighting their limitations in effectively leveraging temporal information from videos, limited tracking ability, and understanding of spatial interactions like physical contact.'*

The problems being tackled in this strand of research are only notionally related to practical robotics at this stage, though additional challenges surely beckon when the theoretical issues finally evolve into embodied AI.

The new paper is titled *Flat-Pack Bench: Evaluating Spatio-Temporal Understanding in Large Vision-Language Models through Furniture Assembly*, and comes from eight authors across Cornell University, Cornell Tech, MBZUAI, and UC Berkeley. The paper is accompanied by a project site.

Method

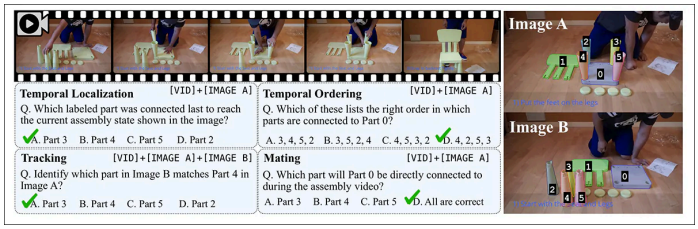
The authors of the new work emphasize the difficulty that AI assistants have in understanding the assembly process through observation, for instance, via the kind of YouTube-style instructional video that many people turn to in order to benefit from community knowledge:



Some of the questions that the flat-pack assembly task provokes, along with the four essential skills necessary to traverse the challenges. Source

They curated a dataset filtered from the earlier-mentioned IKEA-Manuals-at-Work (IMaW) dataset, which features in-the-wild-videos of people assembling IKEA furniture. The revised benchmark trims the original videos to remove text-only instruction cards, with separate key-frame and full-video variants supplied, and also adds manually-annotated visual prompts with segmented furniture parts, to support multiple-choice reasoning tasks.

The benchmark revolves around four question types: *MATE*, determining whether two parts are connected in the final assembly; *TRACK*, requiring models to recover the correct correspondence between shuffled part IDs across segmented frames using the video itself; *TOrd*, evaluating whether models can infer the correct order of connection events; and *TLoc*, testing whether models can identify events occurring immediately before or after the state shown in the visual prompt, requiring temporal localization and reasoning about nearby events.



Examples from the new benchmark, illustrating the four core task types designed to test spatio-temporal reasoning in furniture-assembly videos: Temporal Localiz Temporal Ordering; Tracking; and Mating. Each task combines assembly-video footage with one or more segmentation-labeled visual prompts and a multiple-choi reasoning question.

The templates shown in the schema image above were derived from these four question models.

The authors note also that they added fine-grained part-assembly annotations to each of the original IMaW videos, specifying which parts connect to which other parts – details lacking in the original collection.

Evasion

The questions, the paper notes, needed to be manually curated, since auto-generated questions often give the AI scope to ignore the video and refer to its own trained understanding – a scenario that any regular user of LLMs/VLMs is likely to recognize, since optimization and other mysterious corporate priorities often cause frontier models to ignore submitted information, such as PDFs or images, and rely on their own understanding instead*:

> *'[We] found that auto-generation frequently produced questions that could be answered by ignoring the video and exploiting shortcuts. For example, auto-generated **mating** questions about parts already positioned for connection, or included distractor options with clearly distinct shapes or colors, enabling easy [elimination]. To address this, we curated all questions manually using fixed templates.*
>
> *'Annotators were provided the full assembly video, segmentation-labeled frames for visual prompts, the question templates, and detailed guidelines for avoiding shortcuts based on static cues from the visual prompt.'*

The finished benchmark comprises 602 multiple-choice questions across 50 varying furniture assembly videos.

Data and Tests

Models evaluated for the testing round were the aforementioned ChatGPT and Gemini variants, as well as Video-LLaVA; LLaVA-NeXT-Vid; LLaVA-OneVision; LLaVA-Video; Qwen 2.5/Qwen 3-VL; InternVL3; ArrowRL; PerceptionLM; and Video-Refer.

GenS was used to choose question-relevant frames in long videos for the base Gemini 2.5 Pro model, and most models were tested in a one-shot context under greedy decoding (unsupported in GPT-5, however).

Three prompt formats were devised for the benchmark: the *mixed-media* prompt supplied the visual prompt as a separate image alongside the assembly video; the *collage* prompt embedded the visual prompt directly into every video frame as part of a grid layout; and the *concat* prompt prepended the visual prompts to the start of the video.

Both trimmed and key-frame video variants were tested across these formats, in order to measure how strongly prompt structure and temporal compression might affect model performance.

The chance baselines considered for the tests also included 'frequency chance', where the most common option (rather than a truly random option) is chosen.

Human Factor

Human performance was evaluated using participants drawn from computer science programs, ranging from undergraduate to doctoral level. Each participant was shown an assembly video, and the associated visual prompt and multiple-choice question, as well as the task instruction, before choosing an answer.

Three responses were collected per question and resolved through majority voting, while a separate crowd-sourced study was also conducted on a randomly-sampled subset of the benchmark.

Accuracy was used as the metric for the trials:

Model	Rank	Micro Avg.	TOrd	TLoc	Track	Mate
Human Performance	**–**	94.18	93.54	93.20	93.77	97.70
		Chance Baselines				
Random Chance	**–**	26.41	25.00	25.00	25.49	33.33
Frequency Chance	**–**	26.74	27.74	30.10	26.46	36.78
		Proprietary Models				
GPT-5	**1**	37.71	40.65	53.40	25.68	49.43
Gemini 2.5 Pro	**2**	33.72	40.65	44.66	23.35	39.08
Gemini 3.1 Pro	**3**	32.89	34.84	43.69	21.79	49.43
Gemini 2.5 Flash	**4**	31.06	31.61	41.75	23.35	40.23
Gemini 2.5 Pro + GenS	**5**	25.58	33.55	32.04	13.23	40.23
		Open Models				
Video-LLaVA-7B	**26**	23.75	21.29	35.92	10.89	51.72
InternVL3-14B	**5**	37.71	42.58	21.36	37.74	48.28
InternVL3-38B	**12**	36.05	42.58	37.86	25.68	52.87
InternVL3-78B	**1**	41.03	43.87	39.81	42.02	34.48
Qwen2.5-VL-7B	**22**	30.23	27.10	18.45	33.07	41.38
Qwen2.5-VL-32B	**13**	35.88	34.84	29.13	33.07	54.02
Qwen2.5-VL-72B	**2**	40.37	41.29	30.10	45.14	36.78
Qwen3-VL-4B	**11**	36.54	34.19	33.01	32.68	56.32
Qwen3-VL-4B-Think	**9**	37.21	31.61	25.24	37.74	59.77
Qwen3-VL-8B	**15**	33.72	36.13	30.10	33.85	33.33
Qwen3-VL-8B-Think	**17**	31.73	34.19	33.01	25.29	44.83
Qwen3-VL-32B	**6**	37.71	38.71	46.60	31.91	42.53
Qwen3-VL-32B-Think	**3**	40.03	38.71	22.33	45.53	47.13
Qwen3-VL-30B-A3B	**10**	36.71	30.32	22.33	42.02	49.43
Qwen3-VL-235B-A22B	**8**	37.21	37.42	25.24	39.69	43.68
LLaVA-NeXT-Vid-7B	**25**	25.08	33.55	24.27	16.73	35.63
LLaVA-NeXT-Vid-34B	**21**	30.40	30.32	24.27	32.68	31.03
LlaVA-OneVision-7B	**16**	32.89	26.45	30.10	34.24	43.68
LlaVA-OneVision-72B	**4**	38.37	35.48	25.24	38.91	57.47
LLaVA-Video-7B	**19**	30.73	30.97	24.27	25.68	52.87
LLaVA-Video-72B	**7**	37.54	36.77	27.18	35.80	56.32
Perception-LM-1B	**24**	27.74	28.39	26.21	25.29	35.63
Perception-LM-3B	**18**	31.40	28.39	32.04	29.96	40.23
Perception-LM-8B	**14**	35.38	26.45	26.21	44.75	34.48
VideoRefer	**23**	28.57	32.90	30.10	17.51	51.72
ArrowRL-7B	**20**	30.56	30.97	24.27	29.18	41.38

Performance results on FLAT-PACK BENCH, comparing proprietary and open multimodal models across Temporal Ordering (TOrd), Temporal Localization (TLoc), Tracking, and Mating tasks, with human performance remaining far ahead of all tested systems despite modest gains among larger frontier models.

As seen in the initial tests (image above), humans scored >90% in all categories of questions, with 80% unanimity, suggesting, the paper asserts. that the propositions are well-framed and unambiguous.

GPT-5 and Gemini 2.5/3.1 Pro struggled on the dataset, achieving only modest improvements over the chance baseline, and remained far below human performance. Using GenS to select question-relevant frames did not improve Gemini 2.5 Pro's results, causing the authors to conclude that proprietary LVLMs struggle with the task of spatio-temporal understanding required by the benchmark.

Among open systems, the strongest results came from the InternVL3 and Qwen families, though performance across the category varied sharply, with several models barely outperforming chance; and specialized systems, including PerceptionLM and VideoRefer, also struggled on the benchmark's complex assembly tasks, with human participants remaining significantly ahead in every model category.

The researchers also tested two chain-of-thought prompting strategies against the paper's standard prompting setup. *Zero-shot Chain-of-Thought* prompting asked models to explain their answers step-by-step, while *Self-consistency with Chain-of-Thought* generated five candidate responses before selecting a final answer through majority voting. However, neither improved results on the Flat Pack Bench dataset, with both approaches scoring below the benchmark's default prompting configuration.

Cheat Code

To test whether LVLMs were actually learning from the assembly videos, or merely exploiting static visual cues, the researchers created an image-only version of the benchmark, which omitted the video entirely, retaining only the question-text and visual prompts.

Human performance collapsed by more than 50% under these conditions, showing that the tasks genuinely required temporal understanding of the assembly process. The models, however, degraded far less severely, with some tasks remaining stable or even improving without video input.

This indicates, the paper suggests, that many LVLMs were not meaningfully using the temporal information in the videos *at all*, instead relying on image-based shortcuts and commonsense assumptions to infer plausible answers*:

	Micro Avg.	TORD	TLOC	TRACK	MATE
Video+Image	40.20	40.65	30.10	45.53	35.63
Image-only	31.40	40.65	34.95	21.01	41.38
Δ	–8.80	0.00	+4.85	–24.51	+5.75
Shuffled Parts	29.96 ± 0.69	35.70 ± 3.55	36.25 ± 4.04	20.62 ± 1.68	39.85 ± 2.65
Δ	–10.24	–4.95	+6.15	–24.90	+4.21
Human Perf.	42.69	30.32	45.63	48.24	44.82

Performance of the LVLM on the image-only version of Flat-Pack Bench, compared against the standard video-plus-image setup, with additional results after shuffling part IDs to test whether models were exploiting label-order shortcuts instead of temporal video understanding.

> *'[The image above] shows the performance of the LVLM on this image-only version, and the change in their performance from the full evaluation, along with human performance.*
>
> *'The sharp drop in human performance (>50%) shows that the questions do require videos to answer.*
>
> *'We also observe that the overall performance of the model drops severely (8.80%), but mostly due the TRACK sub-task. Accuracy on other tasks stays the same or improves, indicating that the LVLM **does not use the video effectively**, while humans use the video to answer.'*

The paper's deeper analysis suggests that the main obstacle is not simple temporal sequencing alone, but failures in object grounding and spatio-temporal reasoning: models often struggled to keep track of visually-similar furniture parts across motion, camera shifts, and scene changes, even when they appeared to identify the broader assembly process correctly.

Further experiments involved setting a tool-laden agentic AI loose on the task, and this 'performed poorly' according to the authors – though it was able to correctly answer an additional 11.48% of the questions missed by the other approaches.

Conclusion

Retaining persistent internalizations of concepts and objects is central to both the human experience of growth and perceptual development, and in individual, often novel tasks for which that development has prepared us.

Computer Vision research already has an ongoing struggle to reacquire and re-recognize objects and people that leave and re-enter the frame. These issues are significantly magnified with the need to constantly alter vision and stance – as is likely to occur in a YouTube instructional video about flat-pack furniture assembly. One can imagine the extent to which the even more jarring POV changes of an egocentric video might further confound AI's attempts at furniture assembly.

* *The authors' original formatting, amended by me as necessary to retain impact under quote-formatting/*

First published Monday, May 25, 2026. Amended Wednesday 27th May 2026 to correct this date attribution (!).

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