Episode Book

EPISODE #6175

Video Models Can't Tell Time: What Motion Benchmarks Reveal

PUBLISHED	21 October 2026 at 14:01 UTC
DURATION	23:02
LISTEN	myweirdprompts.com/episode/video-models-motion-benchmarks-30934/
HOSTS	Corn, Herman, Dorothy, Hilbert
SCRIPT MODEL	anthropic/claude-sonnet-5.5
VOICES	chatterbox-regular (chatterbox-local)
PIPELINE	V5.11
BOOK GENERATED	22 October 2026 at 07:46 UTC

SYNOPSIS

Video-capable AI models can describe a room in detail but stumble when asked whether someone is moving slowly, which way they turned, or what happened in what order. Benchmarks like TOMATO, TimeBlind, FAVOR-Bench, and MotionBench all point the same direction: models are good at what's in the frame and shaky about what happens between frames. This episode digs into why — default sampling of one frame per second, missing time bases, and vision encoders that treat each frame in isolation — plus the unglamorous hybrid answer of pairing language models with pose estimation and gait-speed tools.

TRANSCRIPT



Corn

Ninety-eight percent. That's what people score on a test where two videos show the same room, the same objects, the same everything, and only the order of events differs. The best of twenty-plus models, GPT five and Gemini three Pro among them, got forty-eight.



Herman

Forty-eight on what is basically a coin flip with extra steps.



Corn

Which is the territory Daniel is poking at. He's not asking whether a model can say "a man is in a kitchen." He wants to know whether it can say the man is moving slowly, or describe a room properly, when the video goes in and text comes out. Motion, speed, sequence, layout. Who's actually good at it, open or closed, and what do you reach for if the answer is nobody?



Herman

Short version, the models are good at what's in the frame and shaky about what happens between frames. And the reason is almost embarrassing. For a long time you could answer most video questions without watching the video.



Corn

You looked at one still and guessed.



Herman

A single frame, yes. There's a benchmark called TOMATO, from early twenty twenty-five, that was built specifically to fix that. Fourteen hundred and some questions on direction, rotation, speed and frequency, counting repetitions. Humans got ninety-five percent. The best model they tested, a Qwen two VL at seventy-two billion parameters, got thirty-eight, and GPT four o was right behind it.



Corn

Thirty-eight. I've seen sloths do better on a pub quiz, and we don't hold a pen.



Herman

Here's the clever bit. They shuffled the frames. On the older benchmarks, shuffling barely mattered, under eight percent change on average, which tells you the model never used the order at all. On TOMATO the average sensitivity was thirty-four percent. GPT four o dropped from about thirty-eight to about twenty-six when the sixteen frames were scrambled.



Corn

So the old tests were measuring "can you describe a photo," and everybody was calling it video understanding.



Herman

Mostly, yes. A single random frame on TOMATO averaged around twenty-one percent. Sixteen real frames got you the thirty-eight. The gap between those two numbers is the only part that's actually watching.



Corn

Okay, but that's a test problem. Is it also a model problem? Or is it that nobody's feeding the model enough frames to see anything move?



Herman

Both, and they feed each other. Gemini's own documentation says the default is one frame per second, and it warns that fast action can lose detail at that rate. MotionBench, which came out of Zhipu and Tsinghua, points out that a lot of models take sixteen to sixty-four frames in total. Stretch that over a five-minute clip and you're sampling about once every five seconds.



Corn

Once every five seconds. A person can go from sitting to leaving the building in that.



Herman

And you can't judge speed from that. Speed is distance over time, and if the time axis is a handful of snapshots with no clock on them... I mean, the model has the pictures, but the gaps between them are just gone.



Corn

Gemini three Pro is supposed to be better at this, though. They showed a golf swing.



Herman

At ten frames per second, yes. Google says it's tuned for fast action above one frame per second. But ten is an example, not the default, and you pay for it in tokens. I haven't seen a clean motion-benchmark score for Gemini three on this, which is the awkward part. Most of the headline numbers are for last generation's models.



Corn

Meaning the benchmarks are the ones that are slow.



Herman

Ha. Fair. But it does change what you can do today. In Gemini's API you can set the frame rate yourself instead of taking one per second, and you can set the media resolution, so you decide whether the tokens go to more frames or sharper frames. For a person walking across a room, I'd take more frames over sharper ones every time.



Corn

Walking is slow motion for a normal person. Why would it need ten frames a second?



Herman

It doesn't need ten. For gait you probably want a few per second so a foot lift and a foot plant aren't five seconds apart... no, that's wrong, one per second already gives you the steps. What it doesn't give you is the swing. For a golf club, or a hand, or a dropped glass, one per second shows you a before and an after and nothing about the middle.



Corn

And "slowly" is a claim about the middle.



Herman

It's a claim about the rate of change, yes, so you need enough samples inside each movement to see it change. One step per sample is a rumour of a step.



Corn

A rumour of a step. Somebody should put that on the benchmark.



Herman

Ha. So the picture I'd trust is this. MotionBench, early twenty twenty-five, best model around fifty-eight percent, most under sixty. FAVOR-Bench, which has eight thousand questions over seventeen hundred videos, the best model on the multiple-choice side was under fifty. Another one, Vinoground, uses pairs of captions that use the same words in a different order, and GPT four o was around fifty percent against ninety for people.



Corn

Same words, different order. "The man sat then stood" versus the reverse.



Herman

That kind of thing. And the TimeBlind result I opened with is the same trick done with video. Two clips, identical scene, different timing. The authors' conclusion is that the models lean on static cues, because the static cues are right most of the time.



Corn

Correct most of the time is a nasty trap. You get good scores and learn the wrong lesson.



Herman

Let me make TimeBlind concrete, because the design is the neat part. Take one scene and shoot it two ways. In one, a hand picks up the cup and then puts down the phone. In the other, the phone goes down first. Every object, every colour, every background pixel is identical. The only thing that differs is the order and the timing, and the question depends on exactly that.



Corn

So a model that just inventories the scene gets a coin flip, by construction.



Herman

By construction. Six hundred of those pairs, two thousand four hundred video and question combinations. And humans got ninety-eight percent, which tells you the questions are not hard, they're just about time.



Corn

What would you guess is going on inside, Herman? Is the model simply not given the time information, or is it given it and ignores it?



Herman

Honestly, I think it's both, and I can't tell you the split. The vision encoder turns each frame into tokens more or less on its own, and the language model gets a long row of them. Whatever "before" and "after" mean has to be rebuilt from the position in that row, and that's a weak signal next to a vivid picture of a cup. MotionBench's authors proposed fusing neighbouring frames earlier, inside the encoder, rather than after it, for that reason.



Corn

So the order is there in the sequence, but it's whispering.



Herman

Quietly enough that a cup wins. There's a related result on the big general benchmark, Video-MME version two from this April. Gemini three Pro is the best commercial model on it, and it scores about forty-nine on their harsher scoring where humans get about ninety-one. And the odd finding is that adding subtitles helps, and in some cases pure-visual input does worse. The models lean on words when they can.



Corn

That's a sloth's strategy too. If someone's talking, I don't need to watch.



Herman

Which explains your film opinions. Anyway, I'd be cautious reading any single leaderboard. There's a newer paper that splits models by how they work out order. Some read it off positional encoding, Molmo two is the example, and some read it off the actual pixels, Qwen three VL is the example. Two models with the same overall score can fail on completely different clips.



Corn

Same score, different blind spots. Somebody paid for that number.



Herman

Daniel's two examples are good probes for this, by the way. Slowly and the room are both things a person answers from a few seconds of watching. Would you trust a model to do either today, Corn?



Corn

The room, maybe, for a rough description. Chairs, table, window on the left. Slowly, no. I'd want to see it fail first.



Herman

That's roughly what the numbers say, as it happens.



Corn

Well, that's the motion benchmarks. But some tests are stranger than that, aren't they? Hang on, is that my phone... Mum?

D

Dorothy

Hello sweetheart, it's your mother. Time for my grim fact, dear.

D

Dorothy

In two thousand and twenty five, some researchers made videos out of nothing but flickering noise, where the only clue was the order the frames came in. People watching got it right more than ninety eight times out of a hundred, and the very best computer models got none of them right at all, dear. Not one.

D

Dorothy

Anyway. I've got a casserole in and your laundry's still here. When are you coming to see me?



Herman

Mum! Hello, lovely to hear you. We're in the middle of a recording, but we'll ring you after.



Corn

Speed, then, since that's Daniel's own example: is this person moving slowly.



Herman

Speed is where it gets bleakest. There's a paper from this spring on estimating playback speed, so slow motion against real time against sped-up, over a hundred and eleven videos with verified speeds. Gemini two point five, with the sampling raised to eight frames per second, correlated with the truth at about point four. A human expert was point eight eight.



Corn

Point four is "somewhat related."



Herman

And their own small purpose-built model got about point seven four. Smaller, trained for the one job, and it beat Gemini and a plain optical-flow baseline too. Now, that's playback speed, not a person walking at a metre and a half a second. I haven't seen a study that tests a frontier model on real walking speed in metres per second.



Corn

Which is Daniel's actual example. And honestly I'd expect a model to say "slowly" if the person looks hunched and the room is dim, and "briskly" if they're wearing trainers.



Herman

That's the single-frame shortcut again, yes. It reads the scene. A tired-looking man gets "slowly." And there's a deeper issue, which is the model has no clock and no ruler. Slowly compared with what? You need a metric scale, how many metres is that doorway, and a time base, how many seconds between these frames.



Corn

Which a pose estimator has.



Herman

Pose estimation plus a known distance is the unglamorous answer, and it works. In a small study of twenty-four adults over sixty-five, pose-based gait speed matched manual stopwatch timing, one point five five against one point five six metres per second. A smartphone app had a correlation of about point nine four with video. Small, healthy samples, no huge multi-site trial that I found, but it's the sort of thing a language model does not do on its own.



Corn

So the language model's job is to say what's happening, and something else measures it.



Herman

That's the division of labour I'd bet on. There's a fresh benchmark from June, HumanMoveVQA, about ten thousand questions on how a person moves across a scene from a fixed camera, things like trajectory and which way they turned. The proprietary models show a clear gap there too, and fine-tuning an open model on consistent movement data narrows it. I didn't see numbers I could quote, so take it as direction, not a score.



Corn

Fixed camera is the easy case, though. The moment the camera moves, the model has to separate my motion from the camera's.



Herman

MotionBench has camera motion as its own category for that reason, and it's one of the ones models trip on. A handheld clip of someone walking is two motions glued together. A pose tracker in image space can't untangle them either without help, so don't oversell the tool route. It works on a tripod and a doorway.



Corn

A tripod and a doorway. Ambitious.



Herman

That's the hybrid shape. Which would you rather have, Corn? A model that eyeballs it, or one that calls a gait tool and reports the number?



Corn

The tool. I've been eyeballing my own speed for years and the verdict is "yes, slow."



Herman

Good, because the other half of Daniel's question has a similar answer. Describe the room.



Corn

And I'd guess that's easier. A room is mostly still.



Herman

Partly right. VSI-Bench, "Thinking in Space," was built from walkthrough scans of indoor rooms. Humans averaged about seventy-nine percent. The best model in the original paper, Gemini one point five Pro, got forty-five. Room size was nearly human-level, forty-four against forty-six. But route planning, thirty-six against ninety-six, and the order in which objects appeared, thirty-five against a hundred.



Corn

A hundred. The humans got every one of those.



Herman

Because it's the kind of thing you remember effortlessly. The model sees the chair, sees the lamp, and can't tell you which came first. Those are from the original paper, so I wouldn't claim they're current, and I haven't found fresh VSI numbers for Gemini three or Qwen three VL. A December paper, MMSI-Video-Bench, says the best reasoning model among twenty-five still trailed humans by nearly sixty points.



Corn

What does help?



Herman

Chain-of-thought didn't, in either study. Same with extra three-D cues in the newer one. What did help on VSI-Bench was making the model write down a cognitive map first, a rough grid of where things are, and then answer from the map. It sort of makes the layout explicit rather than hoping it's hiding in the weights.



Corn

So the room answer is the same as the speed answer. Don't ask it to hold the geometry in its head. Make it write the geometry down.



Herman

Or get it from a tool. Which is the part Daniel asked about directly, which methods are best. There are four families. Optical flow, which computes motion per pixel between frames and is cheap, though in the playback-speed paper it lost to the dedicated model. Point trackers, CoTracker three from Meta and Oxford, TAPIR from DeepMind, which follows a chosen spot through the clip, TAPIR faster than real time on a decent GPU. Pose estimation for people. And video-native encoders like V-JEPA two, which trains on over a million hours of video with no labels and gets seventy-seven percent on Something-Something v two, a motion-heavy test.



Corn

Meta also released a pile of human-labelled data for this, didn't they?



Herman

Perception LM, yes, millions of fine-grained video questions plus captions that say where and when, not just what. I can't give you reliable per-model scores from it, so I'll only say the direction matches: spend on labelled motion, get motion.



Corn

And then a language model sits on top and calls them.



Herman

Yes, though the evidence there is thin. One training-free method, MotionSight, lifts InternVL three at seventy-eight billion on MotionBench by feeding it motion cues. Flow Four Agent uses optical flow to decide which frames to keep, but that's about token cost, not precision. I haven't seen a clean paper from this year that wires a tracker to a model and scores it on a motion benchmark. It's the obvious thing to do, and nobody's published the obvious thing.



Corn

Or the people who did are selling it. Twelve Labs has models that claim to track motion over time and spit out time-segmented output.



Herman

And their performance claims come from their own materials, so I'd treat those as the vendor's word. What I'd trust more is the training route. FAVOR-Bench's authors fine-tuned Qwen two point five VL on their own training set of seventeen thousand videos, and it improved on MotionBench and other benchmarks they hadn't trained for. Motion is learnable. It just wasn't in the diet.



Corn

Qwen three VL went a different route, didn't it? Timestamps.



Herman

It moved to explicit text timestamps, so the model sees something like a time marker next to the frames instead of inferring time from position. That should help on "when" questions. Whether it helps with "how fast," I don't know. Evals ran at two frames per second, which is better than Gemini's default and nowhere near what a hand movement needs.



Hilbert

Yes, that sounds right. Aunt Ludmila told them the same thing in nineteen ninety-two, on the Harwich ferry, and she'd never seen a television with a freeze-frame button before she took the job.



Corn

What was her job?



Hilbert

Deciding whether a lorry was reversing slowly. Down the car ramp. Eleven pounds forty a day and a tin of herring. She did it from one Polaroid, and she did it by licking the corner. A fast lorry tastes of diesel and a slow one tastes of cardboard.



Herman

That's a sampling problem, in a way. One frame, and she's using a second sense to get at what the frame can't say.



Hilbert

The Polaroids cost one pound seventy a shot, and the camera was on a pigeon. Dennis. Paid in sultanas. Hold on, Herman, you're peaking on the last bit, I'll ride it down... anyway, Dennis only flew when the ramp was down, so every picture was taken from above and slightly behind the lorry.



Corn

Which means every lorry looks like it's leaving.



Hilbert

Every lorry was leaving. That was the whole ruling. The insurer accepted her written opinion in court for four years. I was running the weighbridge at Tilbury then and took the tapes over to her, because the slot was too narrow for my claws and I did it with my tongue. She'd write "slow" on the back of the Polaroid and the insurer paid out on that.



Herman

Four years is a long time for a ruling made off a still image.



Hilbert

She lost when the ferry turned up in the Polaroid. Dennis had been photographing it from behind for the entire summer. It was going backwards. Nobody noticed, because the sea was going the same way.



Corn

What did they do instead?



Hilbert

A man called Pettifer got a stopwatch and two traffic cones and put one at each end of the ramp. Eleven days. Cheaper than the Polaroid film.



Herman

Two cones is a known distance and the stopwatch is a clock. That's the whole of the hybrid, a ruler and a clock bolted on to something that can see.



Corn

And Dennis was still taking pictures, so the clock and the ruler would have told them the ferry was drifting in week one instead of week fourteen.



Hilbert

Week fifteen, with the cones. Fine. I'll give you that.



Herman

What interests me from all of this is the failure that doesn't look like failure. The model is right on most of the clips, so it's confident, so you trust it, and then you hit the clip where the answer was in the order of things and the order was never read.



Corn

Which is the thing worth remembering. The score tells you how often the picture was enough. It doesn't tell you whether the model watched.



Herman

A model that is sixty percent right because it read the scene and a model that is sixty percent right because it tracked the motion look identical on a leaderboard. The shuffle test is the cheap way to tell them apart. Scramble the frames. If nothing changes, nothing was watched.



Corn

Then the practical version for Daniel's two examples. For "is this person moving slowly," don't ask the model to feel it. Give it a pose track and a known distance, and let it read out the number. For "describe the room," have it sketch the layout first, and expect it to be weakest on route and order.



Herman

I'd add that the ranking of open versus closed is muddy. In the tests I've seen this year, the frontier closed models lead on general video benchmarks, Gemini three Pro on top of the commercial ones in one April paper, but the humans are still around ninety and the best model is around fifty on that test's own scoring. On motion specifically, I don't have current numbers that would crown anyone.



Corn

So the ranking is "all of them, badly, for different reasons."



Herman

I'd phrase it kinder. Closed models have the frame budget and the tuning. Open models have the ability to be fine-tuned on motion data, which has been shown to work. Neither one reads time the way you and I do.



Corn

The forward-looking bit, then. The thing I'd watch is whether anyone puts a ruler and a clock inside the model instead of next to it. Two cones and a stopwatch worked for eleven days, and I'd like to see the same thing run at ten frames a second.



Herman

Thanks to Hilbert Flumingtop for producing, as ever.



Corn

This has been My Weird Prompts. You can find every episode at my weird prompts dot com.



Herman

We'll be back soon.